Friedrich-Schiller-Universität Jena
Fakultät für Sozial- und Verhaltenswissenschaften
Institut für Psychologie

Dissertation

Causal Inference in Multilevel Designs

Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena
von Dipl.-Psych. Benjamin Nagengast
geboren am 14.05.1979 in Mainz

Gutachter:

1. Prof. Dr. Rolf Steyer (Friedrich-Schiller-Universität Jena)

2. Prof. Dr. Johannes Hartig (Universität Erfurt)

Tag des Kolloquiums: 11. Juni 2009

Acknowledgments

Many persons supported me during the writing of this thesis, the following list and the mentioned accomplishments are by no means complete and the support often exceeded the areas explicitly included here:

My heartfelt thanks go to Prof. Dr. Rolf Steyer who not only supervised this thesis and stimulated my thinking about causal effects and multilevel designs, in special, and about psychological methodology, in general, but who — at crucial points — encouraged me to pursue a scientific career and supported me in every imaginable way. Special thanks go to Prof. Dr. Johannes Hartig for agreeing to act as external reviewer of this thesis and for the skiing colloquium 2008 on research methods in empirical educational research – I greatly profited from the exchanges and discussions at this meeting.

My parents Silvia and Hans-Joachim Nagengast supported me before, throughout and after my studies, without their love and support, I would not have become the person I am. My brother Arne spurned my ambition to finish this thesis quickly and in time: Thanks for this extra source of motivation! You can do it too!

I am indebted to Ulf Kröhne whose superior programming skills – at least in my humble eyes – allowed me to pursue the sometimes unduly complex simulation studies in this thesis. Moreover, the cigarette breaks on the balcony were often a source of inspiration and sometimes desperately needed for emotion regulation during the last three years. Christiane Fiege and Norman Rose took upon the burden of reading and commenting on earlier versions of this thesis, spotted many inconsistencies and more than one typo. Their comments improved this thesis significantly, the remaining mistakes are solely my responsibility. My colleagues Tim Lossnitzer, Steffi Pohl, Jan Marten Ihme and Hendryk Böhme provided an intellectually stimulating and supportive environment during the last three years. The administrative and organizational skills of Katrin Schaller and Marcel Bauer made my professional life much easier, as did the indispensable work of my student research assistants Andrea Schubert, Marie-Ann Milde and Remo Kamm.

The Deutsche Bahn AG provided a mobile office in the ICs from Weimar to Essen and vice versa for a very reasonable price and did not cancel the late-night connection on sundays.

My friends Aneka Flamm, Florian Kutzner, Hannes Horn and Henri Weise encouraged me to carry on with this thesis at many different occasions.

Last, but in no way least, Anne gave me the love, strength, support and the reason to finish this thesis quickly: The prospects of finally living together again was a major source of motivation to finish it in time and already puts a smile upon my face!

Benjamin Nagengast
Jena, February 2009

Zusammenfassung

In dieser Arbeit wird die allgemeine Theorie kausaler Effekte (Steyer, Partchev, Kröhne, Nagengast, & Fiege, 2009) auf Mehrebenendesigns zum Vergleich der Wirksamkeit verschiedener Behandlungen angewendet und die Bedingungen für kausale Schlüsse in diesen Designs untersucht. In Mehrebenendesigns werden Behandlungseffekte an Beobachtungseinheiten (z.B. Schülern oder Patienten) untersucht, die selbst wiederum in höhere Einheiten (den sogenannten Clustern, z.B. Klassen, Schulen oder Krankenhäusern) geschachtelt sind. Beispiele für solche Designs finden sich in der empirischen Bildungsforschung, der Evaluation von Gruppeninterventionen, z.B. in der Psychotherapieforschung, in der Soziologie, z.B. bei der Untersuchung von Interventionen, die auf der Ebene von Stadtvierteln ansetzen, und in der medizischen Wirksamkeitsforschung, wenn die Effekte einer Behandlung in mehr als einem Krankenhaus untersucht werden.

Konzeptuell lassen sich zwei prototypische Klassen von Mehrebenendesigns unterscheiden: (1) Designs, in denen die Behandlung auf der Ebene der individuellen Beobachtungseinheit ansetzt und (2) Designs, in denen die Behandlungszuweisung auf der Ebene der Cluster stattfindet. Weiterhin und unabhängig von der vorangehenden Dimension, lassen sich Designs mit expliziter Zuweisung von Beobachtungseinheiten zu Clustern und Designs mit bereits existierenden Clustern unterscheiden. Bisherige Arbeiten zur statistischen Analyse von Mehrebenendesigns beschränken sich weitgehend auf experimentelle Designs mit randomisierter Zuweisung von Beobachtungseinheiten oder Clustern zu den Behandlungsbedingungen und vernachlässigen die Analyse nichtrandomisierter und quasi-experimenteller Designs. Nur eine kleine Zahl von Studien befasst sich explizit mit der kausaler Inferenz in Mehrebenendesigns insbesondere auch in nicht-randomisierten Designs. Die darin vorgestellten theoretischen Ansätze sind jedoch entweder zu allgemein formuliert, zu wenig formalisiert oder auf die Betrachtung von Fallstudien beschränkt, und können daher nicht als allgemeine Theorie für kausale Inferenzen in Mehrebenendesigns dienen. Die vorliegende Arbeit schließt diese Lücke

und entwickelt aufbauend auf der allgemeinen Theorie kausaler Effekte die Grundlagen für kausale Inferenz in Mehrebenendesigns.

Im Vergleich zu einfachen Evaluationsdesigns ergeben sich bei der Identifikation und empirischen Schätzung von Behandlungseffekten in Mehrebenendesigns – neben der Auswahl der relevanten Kovariaten und der Spezifkation des Adjustierungsmodells – zusätzliche konzeptuelle wie auch statistische Herausforderungen. Auf der konzeptuellen Ebene ist zu berücksichtigen, dass sich Behandlungseffekte für die Beobachtungseinheiten unterscheiden können, je nachdem, welchem Cluster diese zugeordnet werden. Das Cluster selber kann die Beziehung zwischen Behandlung und Behandlungsergebnis konfundieren. Weiterhin muss berücksichtigt werden, dass stochastische und regressive Abhängigkeiten zwischen Variablen auf den verschiedenen Ebenen des Designs unterschiedlich ausfallen können. Solche sogenannten Kontexteffekte müssen sowohl bei der Definition kausaler Effekte, als auch bei deren statistischer Analyse gesondert berücksichtigt werden. Weiterhin können Interaktionen und Interferenzen zwischen den Beobachtungseinheiten innerhalb eines Clusters oder zwischen den Behandlungsgruppen innerhalb eines Clusters die Interpretation von Behandlungseffekten gefährden. Bei der Formulierung statistischer Modelle ist zu beachten, dass residuale Effekte der Clustervariablen zu einer Unterschätzung von Standardfehlern und liberalen Signifikanztests führen können.

Wie sich zeigt, kann die allgemeine Theorie kausaler Effekte leicht auf Mehrebenendesigns angewendet werden und bietet einen formalisierten Rahmen, um die besonderen Probleme dieser Designs zu lösen. Durch die Definition von sogenannten wahren Effektvariablen bedingt auf alle potentiell konfundierenden Variablen können konfundierende Effekte der Clustervariablen direkt in der elementaren Definition kausaler Effekte berücksichtigt werden. Der durchschnittliche kausale Behandlungseffekt bleibt dabei als Erwartungswert der wahren Effektvariablen wohldefiniert. Vorläufer der allgemeinen Theorie kausaler Effekte und deren Anwendung auf Mehrebenendesigns sind als Spezialfälle in der allgemeinen Theorie enthalten. Die explizite Grundierung der Theorie in einem Einzelversuch, der das Zufallsexperiment des empirischen Phänomens repräsentiert, auf dass sich alle Inferenzen beziehen, erlaubt es zudem die Relevanz von Interferenzen zwischen Beobachtungseinheiten für die Definition kausaler Effekte in verschiedenen Designtypen studieren. Für Designs mit Behandlungszuweisung auf Ebene der Cluster zeigt sich dabei, dass solche Interferenzen nur in Designs mit Zuweisung von Individuen zu Clustern die Validität von Effektdefinitionen gefährden können,

aber auch nur dann, wenn diese Interferenzeffekte nicht vollständig durch Kovariaten erfasst werden. In Designs mit bereits existierenden Clustern sind Interferenzeffekte generell unproblematisch, wenn die Behandlungszuweisung auf der Ebene der Cluster stattfindet. Auch in Designs mit Behandlungszuweisung auf der Ebene der Beobachtungseinheiten sind Interferenzen zwischen behandelten und nicht-behandelten Beobachtungseinheiten innerhalb eines Clusters unproblematisch, solange sie als Funktion der Clustervariable aufgefasst werden können. Unabhängig von der Art des Designs sind valide Schlüsse aus Stichproben an die Voraussetzung der Wiederholung kausalstabiler Einzelversuche geknüpft und die Generalisierbarkeit von Befunden ohne weitere Annahmen auf das durch den Einzelversuch und die entsprechenden Verteilungen und Parameter repräsentierte Design beschränkt.

Auf der Grundlage der allgemeinen Theorie kausaler Effekte wurden im folgenden generalisierte Kovarianzanalysen (ANCOVA) zur Schätzung durchschnittlicher kausaler Effekte für bedingt-randomisierte und quasi-experimentelle Designs mit Behandlungszuweisung auf individueller Ebene und auf der Ebene des Clusters dargestellt. Dabei wurden diese Verfahren zunächst für allgemeine bedingte Effektfunktionen entwickelt und dann für lineare Effektfunktionen spezifiziert. Herkömmliche Ansätze der Kovarianzanalyse für Mehrebenenmodelle werden erweitert, indem einerseits Interaktionen zwischen der Behandlungsvariablen und den Kovariaten zugelassen und dabei auch Kontexteffekte berücksichtigt werden, andererseits der durchschnittlichen kausale Effekt eindeutig identifiziert wird. Die Implementierung der generalisierten ANCOVA in verschiedenen statistischen Verfahren wurde in zwei separaten Simulationsstudien für Designs mit Behandlungszuweisung auf Individuuen- und auf Clusterebene getestet und die Modelle auf Datenbeispiele angewandt. Dabei zeigte sich, dass sowohl die Mehrebenenstruktur der Daten als auch die Stochastizität der Prädiktoren bei der Bestimmung von Standardfehlern und bei Signifikanztests berücksichtigt werden muss. In Designs mit Behandlungszuweisung auf der Clusterebene schätzten nur solche Verfahren den durchschnittlichen kausalen Effekt erwartungstreu, die berücksichtigten, dass die empirischen Mittelwerte der Kovariaten innerhalb der Cluster die bedingten Erwartungswerte der Kovariaten nur fehlerbehaftet messen. Statistische Verfahren, die dies nicht berücksichtigten, zeigten unter bestimmten Bedingungen einen Bias in der Parameterschätzung. Das vielversprechendste Verfahren in beiden Simulationen, die Implementierung des hierarchichen linearen Regression als Mehrebenenstrukturgleichungsmodell in `Mplus`, wies jedoch unter realistischen Parameterkonstellationen

teilweise Konvergenzprobleme auf und führte teilweise zu leichten Verschätzungen der Standardfehler, so dass es nicht vorbehaltlos für den Einsatz in der Praxis empfohlen werden kann.

In der abschließenden Diskussion wird ausführlich auf den Geltungsbereich der allgemeinen Theorie kausaler Effekte eingegangen. Außerdem werden die Vor- und Nachteile der generalisierten ANCOVA für Mehrebenendesigns und ihrer Implementierung in verschiedenen statistischen Modellen diskutiert, kritische Annahmen expliziert und Alternativverfahren kurz vorgestellt. Abschließend werden die noch offenen Fragen für kausale Schlüsse in Mehrebenendesigns kurz vorgestellt.

Abstract

The general theory of causal effects (Steyer et al., 2009) is used to develop a theory of causal inference for multilevel designs - i.e., for designs in which the effects of treatments are evaluated on units nested within clusters - that extends and consolidates previous approaches. Two multilevel causality spaces for different classes of multilevel designs are used to define true-effect variables, average causal effects, conditional causal effects and prima-facie effects. Unbiasedness, as the weakest condition under which average and conditional causal effects are identified, and its sufficient conditions are outlined. Next, stability assumptions for causal inference in multilevel designs are discussed in relation to the general theory of causal effects and a taxonomy of multilevel designs is introduced. Building upon this theoretical framework, the generalized analysis of covariance (ANCOVA), that extends the conventional multilevel ANCOVA by identifying the average causal effect in the presence of interactions, is developed for non-randomized multilevel designs with treatment assignment at unit- and at the cluster-level. Two simulation studies tested several statistical implementations of the generalized ANCOVAs. The results showed that contextual effects have to be taken into account in the specification of adjustment models, that predictors have to be modeled as stochastic to obtain correct standard errors of the average causal effects and that the unreliability of the empirical cluster means has to be accounted for in designs with treatment assignment at the cluster-level. The statistical methods studied in the simulations were applied to two empirical examples from educational research to demonstrate the implementations in practice. Finally, the scope of the general theory of causal effects, the advantages and disadvantages of the generalized ANCOVA and alternative adjustment methods are discussed and an overview of further research needs is given.

Contents

List of Figures

List of Tables

1 Introduction

This thesis is concerned with causal inference in multilevel designs — more specifically with causal inference in multilevel between-group designs. In such designs, the effects of a treatment — for example a new teaching method, a psychotherapy or a medical procedure — on an outcome compared to a control condition — for example, a conventional teaching method, an established psychotherapy or no treatment at all — are evaluated on a group of subjects who are nested within higher hierarchical units — for example on students nested within classrooms, on patients nested within hospitals or treatment groups or on inhabitants of different neighborhoods. Throughout this thesis, we will refer to the entities at the lowest level of nesting as *units* and to the structures they are nested in as *clusters*. In contrast, we will use the terms *treatment* or *control conditions* and *groups* respectively when referring to the intervention that is being evaluated. We will focus exclusively on designs in which one or more clearly defined treatments are compared to clearly defined control conditions and all of them can be implemented — at least theoretically — in more than one cluster at the same time. Studies in which the effect of being in a specific cluster is of interest are not covered in this thesis. However, the theoretical framework is, in principle, flexible enough to deal with the peculiarities of such studies.

The "fundamental problem of causal inference" (Holland, 1986, p. 947) is exacerbated in multilevel between-group designs: Each unit can only be observed in one specific cluster and assigned to one treatment condition (Gitelman, 2005). In this thesis, we will study if, when and how the nested structure of multilevel designs has to be taken into account in causal inferences about the treatment effects and the ensuing statistical analysis. In contrast to singlelevel between-group designs for which several accounts of causal inference in different research traditions exist (e.g., Shadish, Cook, & Campbell, 2002; Steyer et al., 2009), the conditions for causal inference in multilevel designs have not been developed with the same rigor (Schafer & Kang, 2008). In fact, multilevel designs have received at best marginal mention in current textbooks

on experimental design (e.g., Hinkelmann & Kempthorne, 2005; Maxwell & Delaney, 2004; Montgomery, 2001; Shadish et al., 2002) although they are important in various areas such as health research (e.g., Donner & Klar, 2000; Murray, 1998; Turpin & Sinacore, 1991), education (e.g., Raudenbush, 2003) and public policy analysis (e.g., Sobel, 2006). This thesis seeks to fill the gap in the literature and to develop the theoretical prerequisites and general statistical models for causal inference in multilevel designs. In this chapter, we will informally introduce different multilevel between-group designs, briefly review the literature on the statistical analysis of these designs, which focuses heavily on randomized designs, and then discuss the merits and shortcomings of existing accounts of causal inference in multilevel designs. The chapter closes with an overview and outlook on the structure and contents of this thesis.

1.1 Multilevel Designs

We start with a brief review of different multilevel designs for which we intend to develop a framework for causal inference. The different designs will only be informally introduced here. A more thorough discussion and taxonomy of multilevel designs in light of an explicit theory of causality will be given in Chapter 3.

Conventionally, between-group multilevel designs are characterized by two properties: (1) The level at which treatment assignment takes place and (2) the treatment assignment mechanism (Plewis & Hurry, 1998; Seltzer, 2004). In this very coarse introduction of multilevel designs, we will place our focus on the level of treatment assignment and restrict our discussion to designs with one level of nesting, i.e., to designs in which units are nested within clusters. We will broadly distinguish between designs with treatment assignment at the unit-level (e.g., at the level of individual students or patients) and designs with treatment assignment at the cluster-level (e.g., at the level of classrooms, hospitals or neighborhoods, see also, Moerbeek, van Breukelen, & Berger, 2000; Plewis & Hurry, 1998, for similar frameworks). Designs with additional levels of nesting are discussed by Schochet (2008) for educational examples. Although a generalization to more complex designs would be desirable, the complexities involved are beyond the scope of this thesis. In this introduction, we will only coarsely differentiate between (1) randomized designs, in which each unit of assignment has the same probability of being assigned to the treatment conditions and (2)

non-randomized designs, in which treatment assignment probabilities differ between the units of assignment. A more thorough discussion of the distinction between experimental and quasi-experimental designs and the specific challenges to causal inference is delayed to Chapter 3.

In multilevel designs with *treatment assignment at the unit-level*, each unit within a cluster is assigned individually to either the treatment or the control condition. Within each cluster (e.g., classrooms, schools, hospitals, neighborhoods) a small experiment or quasi-experiment is conducted. If the assignment probabilities are the same for all clusters and units, the design is referred to as a randomized multisite trial (Raudenbush & Liu, 2000; Turpin & Sinacore, 1991). There are a variety of ways in which treatment assignment probabilities can systematically differ between units and clusters that will be discussed in more detail in Chapter 3. One design type worth mentioning here are blocked randomized designs in which assignment probabilities are equal for all units within a cluster, but can differ between clusters (Bloom, Bos, & Lee, 1999). Examples of multilevel designs with treatment assignment at the unit-level in health care and criminal justice settings are given by Turpin and Sinacore (1991).

In multilevel designs with *treatment assignment at the cluster-level*, clusters (e.g., classrooms, schools, hospitals or neighborhoods) are assigned to treatment conditions as a whole (Donner & Klar, 2000; Murray, 1998). All units within a cluster receive the same treatment and usually the treatment is administered to all units within a cluster at the same time. When clusters are assigned randomly to treatment conditions, the corresponding design is referred to as a cluster randomized trial (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). Like randomized experiments in singlelevel designs, cluster randomized trials guarantee that treatment and control groups are similar on average, even if the clusters differ in composition. In non-randomized designs with treatment assignment at the cluster-level, assignment probabilities can differ between the clusters — reflecting the influence of cluster-level variables such as therapist characteristics, composition of the cluster or geographic location. The reviews by Ukoumunne, Gulliford, Chinn, Sterne, and Burney (1999) and Varnell, Murray, Janega, and Blitstein (2004) include a large number of examples for randomized and non-randomized multilevel designs with treatment assignment at the cluster-level in public health research. The review by Baldwin, Murray, and Shadish (2005) lists examples from psychotherapy research.

A third type of multilevel designs, that will not be covered in detail are studies in

which the effect of being in a cluster is of interest. Such research questions are relevant in the context of value-added modeling and accountability systems in education (e.g., McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004) or hospital profiling (e.g., Wegscheider, 2004). Here, the effect of being taught by a specific teacher or in a specific school or of being treated in a specific hospital, is of interest. The complexities of causal inference for these purposes are discussed extensively by Fiege (2007), Raudenbush and Willms (1995) and Rubin, Stuart, and Zanutto (2004) among others. Contrary to the between-group multilevel designs introduced above, the treatment cannot be thought of as being implementable in more than one cluster at a time and the elements of the treatment itself are sometimes hard to define (Raudenbush & Willms, 1995; Rubin et al., 2004). A detailed discussion of the profiling literature is not within the scope of this thesis, in which we will focus solely on between-group multilevel designs with clearly defined treatment and comparison conditions that could be implemented concurrently at different sites.

1.2 Statistical Inference in Multilevel Designs

There is a large literature dealing with the statistical analysis of between-group multilevel designs, covering designs with treatment assignment at the unit-level as well as designs with treatment assignment at the cluster-level. These papers almost completely fail to base their discussion upon an explicit theory of causality and are mostly restricted to the analysis of randomized designs. We will briefly review the most important findings before we turn to the existing discussions of causal inference in multilevel designs.

There is a broad literature on the analysis of cluster-randomized studies dating back to Walsh's (1947) development of correction formulae for standard errors of mean comparisons. The consensus in the literature is that the clustered data structure needs to be taken into account in statistical analysis (e.g., Donner & Klar, 2000; Feng, Diehr, Peterson, & McLerran, 2001; Murray, 1998, 2001; Murray, Varnell, & Blitstein, 2004; Hedges, 2007a). Ignoring the clustered data structure leads to spuriously small standard errors and inflated type-1-error rates even for very small effects of clustering (e.g., Baldwin et al., 2005). Multilevel models such as the hierarchical linear model (e.g., Goldstein, 1999; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) offer the largest

flexibility in accounting for the clustered structure, because they can naturally handle unequal cluster sizes and the inclusion of covariates (Moerbeek, van Breukelen, & Berger, 2003).

Although there is a strong consensus in the statistical literature about the importance of accounting for the clustered data-structure with appropriate statistical methods, researchers do not consistently apply these methods in practice: Literature reviews in public health (Varnell et al., 2004; Ukoumunne et al., 1999) and in psychotherapy research (Baldwin et al., 2005) suggest that a considerable number of studies is still analyzed with non-appropriate techniques – putting the analyses in danger of spuriously interpreting non-significant treatment effects as statistically significant. Varnell et al. (2004) reported that just 54.2% of the studies in their sample employed only appropriate methods for the analysis of group randomized trials. Baldwin et al. (2005) demonstrated that appropriately accounting for clustering significantly decreased the number of statistically significant results in published evaluations of group psychotherapy.

Power analysis and optimal cost-effective design strategies are well-understood for randomized multilevel designs with two levels of nesting (Hayes & Bennet, 1999; Moerbeek, 2008; Moerbeek et al., 2000; Raudenbush, 1997; Raudenbush & Liu, 2000; Schochet, 2008; Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2008). Comparisons of designs with treatment assignment at the unit-level and treatment assignment at the cluster-level indicate that the former yield a higher power to detect treatment effects (Moerbeek, van Breukelen, & Berger, 2001; Schochet, 2008) and are thus to be preferred if their implementation is feasible. Power calculations for designs with more than one level of nesting are discussed by Konstantopoulos (2008) and Schochet (2008). Hedges (2007b) offers a discussion of different methods for computing effect sizes in multilevel designs. The intraclass correlation coefficient (*ICC*) and the related variance inflation factor (*VIF*, Kish, 1965) reflect the amount to which the clustering affects statistical inferences and power (see also Section 2.3). Empirical *ICC* values needed for a-priori power calculations are available for applications in public health (Gulliford, Ukoumunne, & Chinn, 1999), in medical research (Turner, Thompson, & Spiegelhalter, 2005) and in educational studies (Hedges & Hedberg, 2007; Schochet, 2008).

Reviews of analytical strategies for multilevel designs discuss the inclusion of covariates or pretests in analytical methods only with reference to the reduction the residual error variance and the increased power to detect treatment effects (e.g., Bloom et al.,

1999; Bloom, Richburg-Hayes, & Black, 2007; Murray, Van Horn, Hawkins, & Arthur, 2006; Plewis & Hurry, 1998). The importance of treatment-covariate interactions is rarely mentioned (see, Pituch, 2001; Plewis & Hurry, 1998; Seltzer, 2004, for notable exceptions) and the appropriate analysis of average treatment effects in their presence has not been thoroughly studied (see also, Gitelman, 2005; VanderWeele, 2008). Optimal design strategies for randomized trials with covariates are given by Moerbeek et al. (2001) who also discuss stratified assignment as a means to reduce dependencies between treatment assignment and covariates. Raudenbush, Martinez, and Spybrook (2007) discuss the relative efficiency of true randomization versus stratified allocation of clusters and the related analytical methods for designs with treatment assignment at the cluster-level.

Mediation in multilevel designs is discussed in detail by Krull and MacKinnon (2001) and Raudenbush and Sampson (1999). Pituch, Whittaker, and Stapleton (2005) compare different methods for tests of mediational effects in multisite experiments, Pituch, Stapleton, and Kang (2006) extend this analysis to tests of mediation in cluster-randomized designs. D. J. Bauer, Preacher, and Gil (2006) cover moderated mediation in multilevel models. However, this literature is devoid of references to the problem of interpreting mediational effects as causal effects and the necessary assumptions (Sobel, 2008). Non-compliance in cluster-randomized studies is analyzed by Frangakis, Rubin, and Zhou (2002) and Jo, Asparouhov, Muthén, Ialongo, and Brown (2008) in a principal stratification framework (Frangakis & Rubin, 2002).

In general, the literature on the statistical analysis of multilevel designs is astonishingly little concerned with the foundation of their inferences in a theoretical account of causality. Apart from casual discussions (e.g., D. J. Bauer, Sterba, & Hallfors, 2008; Bingenheimer & Raudenbush, 2004; Seltzer, 2004) causal inference is barely mentioned as a goal of statistical analyses. Instead, discussions of the intricacies of statistical models are given comparatively much room. In the next section, we will review the exceptions to this rule — the approaches to an explicit account of causality in multilevel designs.

1.3 Causal Inference in Multilevel Designs

Theories of causal effects in between-group designs (Neyman, 1923/1990; Rubin, 1974, 1977, 1978; Steyer et al., 2009) address the fundamental problem of causal inference (Holland, 1986) — the impossibility to expose a unit to more than one treatment condition at the same time and observe the corresponding outcomes to obtain a direct measure of the individual treatment effect. To circumvent this problem, these theories define the average causal treatment effect as the average of the individual treatment effects and specify the conditions under which this theoretical quantity can be identified with empirically estimable quantities. As a final step, practical estimation in statistical sampling models is considered (Manski, 1995; Steyer et al., 2009).

The theory of individual and average causal effects (Rubin, 1974, 1977, 1978) and its generalizations (Neyman, 1923/1990; Steyer et al., 2009) are the most popular frameworks for discussing causal inference in singlelevel between-group experiments and quasi-experiments. The core concept in these theoretical accounts is the average causal treatment effect. In randomized experiments, the average causal effect can be identified with the difference of the means of the outcome variable in the treatment and the control group. In quasi-experimental designs, one out of several sufficient conditions (with strong ignorability as the most popular, Morgan & Winship, 2007; Rubin, 1977) must be fulfilled in order to identify the average causal effect. Several methods for the estimation of average causal effects have been developed: analysis of covariance (ANCOVA) and its generalizations for models with interactions between treatment and covariates and non-linear effect functions (Steyer et al., 2009), propensity score methods (e.g., Rosenbaum & Rubin, 1983) such as stratification (e.g., Rosenbaum & Rubin, 1984), matching (e.g., Rosenbaum & Rubin, 1985) and weighting (e.g., Hirano & Imbens, 2001) as well as instrumental variable estimators (e.g., Angrist, Imbens, & Rubin, 1996).

The literature on causal inference for multilevel designs is relatively sparse and a comprehensive and coherent framework is lacking. This is regrettable, because the nested structure of multilevel designs poses unique challenges to causal inference: The fundamental problem of causal inference (Holland, 1986) is exacerbated: The outcome of each unit can not only be observed only in one treatment condition, but also only within one cluster (Gitelman, 2005). Contextual effects (Greenland, 1992), aggregation bias (Alker, 1969; Robinson, 1950) and the appropriate representation of the nested

design structure in statistical analyses further complicate causal inferences. Classical textbooks on experimental design and causal inference (e.g., Maxwell & Delaney, 2004; Shadish et al., 2002) mention multilevel designs only marginally (but see, Gelman & Hill, 2007, for a notable exception), are confined to randomized designs and do not embed their discussion in an explicit formal theory of causality. As outlined in the previous section, the literature on statistical analysis of multilevel designs suffers from the same shortcomings: Most authors who discuss the analysis of multilevel experiments and quasi-experiments confine their discussions to designs with randomization either on the level of individuals or on the level of clusters (e.g., Moerbeek et al., 2000, 2001; Moerbeek, van Breukelen, Berger, & Ausems, 2003; Raudenbush & Liu, 2000; Schochet, 2008) or — if they consider quasi-experiments or observational studies explicitly — rely on informal definitions of causality (e.g., Seltzer, 2004). There are some noteworthy exceptions (Gitelman, 2005; Hong & Raudenbush, 2006, 2008; Raudenbush, Hong, & Rowan, 2006; Sobel, 2006; VanderWeele, 2008) that apply the potential-outcome framework (Rubin, 1974, 1977, 1978) to some aspects of multilevel designs, sometimes additionally invoking the theory of acyclical directed graphs (Pearl, 2000). However, a generalization to the more general theory of causal inference introduced by Steyer et al. (2009), including the concepts of true-outcome variables (Steyer, Gabler, von Davier, Nachtigall, & Buhl, 2000; Steyer, Nachtigall, Wüthrich-Martone, & Kraus, 2002; Steyer et al., 2009) and conditions for causal inference other than strong ignorability (Steyer, Gabler, von Davier, & Nachtigall, 2000) is lacking.

1.3.1 Literature Review

The present approaches to causal inference can be broadly classified into two categories: Among the authors who tackle the problem of causal inference in multilevel designs very generally, some try to give broad, rather informal overviews about its prospects (Draper, 1995; Oakes, 2004), while others directly apply the potential-outcome framework to some standard designs (Gitelman, 2005; VanderWeele, 2008). A second group of authors uses the potential-outcome framework (Rubin, 1974, 1977, 1978) to discuss causal inference for specific case studies and narrowly defined problems, e.g., the effect of instructional sequences (Hong & Raudenbush, 2008; Raudenbush et al., 2006), retention policy (Hong & Raudenbush, 2006) or neighborhood interventions (Sobel, 2006). In the following sections, we will briefly introduce these

papers, outline common themes and discuss inconsistencies and shortcomings. Our discussion will start with papers that treat the prospects of causal inference generally (Draper, 1995; Gitelman, 2005; Oakes, 2004; VanderWeele, 2008) and then examine the merits of some applications of the potential outcome framework to case studies (Hong & Raudenbush, 2006, 2008; Raudenbush et al., 2006; Sobel, 2006).

General Approaches

The papers by Draper (1995) and Oakes (2004) try to tackle the problem of causal inference in multilevel designs very generally. Although they suffer from a lack of formalization and rely on informal introductions of central concepts, both papers inspired controversial discussions. Gitelman (2005) and VanderWeele (2008) specifically adapt the potential-outcome framework of Rubin (1974, 1977, 1978) to deal with some of the complexities of multilevel designs. Both papers, however, fall short of providing a truly comprehensive framework of causal inference for multilevel designs.

Draper (1995) was one of the first to discuss causal inference in multilevel models with an explicit reference to the potential-outcome perspective. His paper dealt with a variety of other issues in multilevel modeling (e.g., estimation), but also inspired a lively discussion about the prospects of causal inference in these kinds of models (see, e.g., the responses by Longford, 1995; Raudenbush, 1995). Drawing upon the Campbellian tradition (Campbell & Stanley, 1966; Shadish et al., 2002) of external validity as well as on the potential-outcome framework (Rubin, 1974, 1977, 1978), Draper argued that valid causal inferences in multilevel designs would only be possible in studies with a representative sample of the target population — more specifically in studies that sampled units randomly from a well-defined population — and with a strongly ignorable treatment assignment scheme (Rubin, 1977). Draper vehemently argued for controlled randomized experiments (or at least stratified designs) at different levels of nesting in a multilevel framework to heighten the inferential strength of evaluation studies in multilevel contexts.

Among the responses to Draper (1995), the comments of Raudenbush (1995) are especially noteworthy. Raudenbush criticized Draper's strict advocacy of randomized studies, noting that "we cannot wait for the perfect social science study that randomly selects subjects from a large and well-defined population and then randomly assigns subjects to treatment [. . .] to allow unquestionable and generalizable causal inferences"

(p. 213). Raudenbush noted that representative samples were neither a typical nor a necessary condition for the advancement of knowledge and theory building in the social sciences and useful mainly as a "vision [...] to promote better research practice" (p. 213). Instead, he emphasized the distinction between *statistical* inference and *causal* inference and argued that the question of when and under which assumptions parameters of a multilevel model represented causal entities should be carefully studied for specific designs.

Oakes (2004) discussed the prospects of causal interpretation of neighborhood effects in social epidemiology and their estimation with linear multilevel models. He identified four methodological obstacles that, in his view, precluded the interpretation of neighborhood effects as causal effects: (1) He argued that accounting for the mechanism that produced different neighborhood compositions with a level-1-model would lead to complete separation of neighborhoods with respect to these covariates, i.e., missing overlap of the corresponding covariate distributions. (2) He conjectured that context variables, such as the mean income or health status within a neighborhood, were emergent properties of the neighborhood composition, and as such endogeneous variables (see, Manski, 1995), and controlling for them would violate model assumptions. (3) He argued that the problem of extrapolation over the observed range of covariates would be especially critical for causal inference in multilevel contexts where little overlap in the covariate distribution between clusters could be expected. (4) He referred to violations of the *Stable Unit Treatment Value Assumption* (SUTVA, Rubin, 1977, 1986, 1990), noting that a causal interpretation of neighborhood effects was endangered by disequilibria resulting from relocation of persons to different neighborhoods. As an alternative to observational studies of neighborhood effects, Oakes argued for controlled community trials to investigate the effects of specific interventions at the level of communities.

Oakes (2004) did not clearly distinguish between problems in *design* and *analysis* of studies of neighborhood effects at many points in his discussion (see also, Subramanian, 2004). Additionally and even more damaging, he failed to unambiguously define neighborhood effects in terms of an explicit theory of causality and it remains unclear whether the effects of a specific neighborhood or of certain neighborhood characteristics are implied by his definition. The four main challenges to causal inference in neighborhood designs are either empirically testable (*separation between neighborhoods*, Diez Roux, 2004; Subramanian, 2004), too pessimistic (*endogenous effects*, *SUTVA violations*, Gitelman, 2005; Hong & Raudenbush, 2006; Manski, 1995) or not

specific to multilevel designs, but generally relevant for the analysis of causal effects with linear models (*extrapolation*, Rosenbaum & Rubin, 1983; Rubin, 1973, 1977). In any case, they do not preclude the causal interpretation of treatment effects in multilevel designs in general.

Gitelman (2005) provided a theoretical discussion of causal inference for designs with treatment assignment at the cluster-level. She focused on possible violations of SUTVA (Rubin, 1977, 1986, 1990) and used the theory of directed acyclical graphs (Pearl, 2000) to derive the independence assumptions made in causal inference from group allocation data. Based on these assumptions, she expanded the potential-outcome framework (Rubin, 1974, 1977, 1978) to allow varying potential-outcomes for every treatment-cluster combination. Based on this expansion she defined a "group-allocation, multilevel average" (p. 406) causal effect (*GAMA*). Extending the concept of strong ignorability (Rubin, 1977), she identified this effect with a model parameter in a hierarchical linear model under the assumption of stochastic independence of the cluster and treatment allocation from the potential-outcome variables conditional on subject-level covariates.

Though representing one of the most advanced accounts of causal inference in multilevel designs, Gitelman's (2005) work has major shortcomings: (1) It only covers designs with treatment assignment at the cluster-level, leaving out multilevel designs with treatment assignment at the unit-level. (2) In light of the general theory of causal effects (Steyer et al., 2009), her assumption about the absence of group-dynamic effects is unnecessarily rigid, as we will show in Chapter 3 in more detail. (3) Finally, her identification of the average causal effect with parameters of a hierarchical linear model is flawed and depends implicitly on the absence of interactions between covariates and the treatments or on the covariates having an expected value of zero. In Chapters 4 and 5, we will develop generalized ANCOVAs that identify the average causal effect even in the presence of interactions between treatment and covariates and non-zero expected values of the covariates.

VanderWeele (2008) discussed the relevance of ignorability and stability assumptions in causal inference for neighborhood effects research in social epidemiology. He focused exclusively on designs in which treatments are applied at the level of neighborhoods. In order to handle causal inference for observational studies in these settings, he introduced the notion of the *Neighborhood-Level Stable Treatment Unit Value Assumption* (NL-SUTVA) and extended the definition of strong ignorability to include

covariates at the neighborhood-level. He then showed how the average treatment effect could be identified with the regression coefficient of the treatment indicator in a linear multilevel model provided all relevant covariates were included in the model, no treatment-covariate interactions were present and the functional form of the regression were correctly specified. He also argued that covariates at the individual-level would only have to be included in the model, if aspects of the individual-level distribution of the covariate other than the corresponding neighborhoods means influenced the treatment assignment (see also Chapter 5. Contrary to Gitelman (2005), VanderWeele did not allow potential-outcomes to vary between neighborhoods. He also did not discuss how to identify average causal effects in the presence of interactions and how to model these effects in applications. We will return to VanderWeele's stability assumption in Chpater 3 and show how it can be reconciled with Gitelman's assumption.

Case Studies

The papers by Hong and Raudenbush (2006, 2008), Raudenbush et al. (2006) and Sobel (2006) are case studies that apply certain aspects of the potential-outcome framework for causal inference (Rubin, 1974, 1977, 1978) to specific research questions and applications in multilevel contexts. While highly sophisticated in details, they are always restricted to the specific applications and do not generalize to multilevel designs in general. We will review these studies in the remainder of this section.

Hong and Raudenbush (2006) discussed causal inference in a case study evaluating the effects of kindergarten retention and retention policy. They extended the framework for causal inference (Rubin, 1974, 1977, 1978) by loosening SUTVA (Rubin, 1977, 1986, 1990) for their designs with treatment assignment at the unit-level and allowing potential-outcomes to vary with functions of the treatment assignment vector in a cluster. In their specific example, individual causal effects were allowed to vary by the retention policy of the school which was conceived of as a function of the individual retention probabilities of the students within that school. They explicitly confined their discussion of causal effects to the current allocation of students to schools (see, Fiege, 2007, for a similar point regarding causal inference in school comparisons) and did not generalize their framework to include potential-outcomes for every student-school combination. Based on these restrictions, potential-outcomes were identified for three subgroups of children that were either always, sometimes — under a high retention po-

licy — or never at risk of being retained in kindergarten. In a next step, they identified some of the average causal effects with empirically estimable quantities under the assumption of strong ignorability conditional on student- and school-covariates. Finally, they empirically estimated these average effects using propensity scores to capture the effects of multiple covariates.

In two closely related papers, Hong and Raudenbush (2008) and Raudenbush et al. (2006) applied Rubin's framework for causal inference (1974, 1977, 1978) to the evaluation of effects of instructional sequences over more than one academical year. They studied the effects of a treatment at the classroom-level – intensive versus standard instruction in mathematics. Drawing upon concepts developed in epidemiology (Robins, 1987, 2000; Robins, Hérnan, & Brumback, 2000) for time-constant and time-varying covariates, they developed the notion of "sequential strong ignorability" (Hong & Raudenbush, 2008, p. 341) to describe designs where randomization at each timepoint is conditional upon all observed covariates (time-variant and time-invariant) and prior treatments received by a person. They defined four types of average causal effects of instructional sequences for their case study and showed how they could be identified as deflections from an individual linear growth trajectory over time. In order to consistently estimate the effects of intensive instruction over three years of schooling, they used inverse probability-of-treatment weighting based on an estimated propensity score and pseudolikelihood estimation. Like Hong and Raudenbush (2006), they confined their inferences to intact schools, defining potential-outcomes for students only with respect to the actual school and class assignment.

Sobel (2006) developed a framework for causal inference for randomized studies of neighborhood effects that takes interactions between members of a common neighborhood into account and thus relaxes SUTVA (Rubin, 1977, 1990). He developed his framework in the context of a complex evaluation study of a neighborhood intervention program in which inhabitants of disadvantaged city areas were randomly assigned to receive assistance to relocate to more affluent areas. Sobel placed his derivations within the framework of causal effects analysis with instrumental variables introduced by Angrist et al. (1996). By taking all possible allocations of units to treatments explicitly into account, he defined plausible average causal effects to cover the specificities of the intervention design. He then demonstrated that the published estimates of the average causal effect rely on the assumption of no spillover effects between participants who received the treatment and participants who did not receive the treatment and are biased

when such effects are present within one neighborhood. He showed, that in order to estimate a meaningful average causal effect, a comparison with a control neighborhood would be necessary where no opportunities to move to other neighborhoods existed. Consequently, he argued for group randomized designs where neighborhoods were randomly assigned to either receive the assignment scheme of the intervention program or no treatment at all and outlined the necessary stability assumption in these designs.

1.3.2 Common Themes and Challenges

The state of the literature of causal inference in multilevel models can be summarized as follows: A common well-defined framework for causal inference in these designs is still lacking. Although there are some encouraging attempts to apply the Rubin causal model (Rubin, 1974, 1977, 1978) to multilevel designs, they ultimately fall short of providing an encompassing theory of causal inference for multilevel designs. They are either confined to a single design type (e.g., Gitelman, 2005; VanderWeele, 2008) or deal with special problems or applications (Hong & Raudenbush, 2006, 2008; Sobel, 2006). Additionally, no consensus seems to have been reached about the necessary and sufficient conditions for causal inference in multilevel designs. The prevalence of ill-defined concepts is apparent in some of the papers that are restricted to informal definitions of causal effects (Draper, 1995; Oakes, 2004), or do not specify the interesting concepts unambiguously (Gitelman, 2005; VanderWeele, 2008). As mentioned above, accounts of statistical analyses of experimental and quasi-experimental multilevel designs often fail to ground their discussion in an explicit theory of causal effects (e.g., Seltzer, 2004).

There are a number of common themes and shortcomings of the present approaches that should be highlighted and summarized: Possible violations of the *Stable Unit Treatment Value Assumption* (SUTVA, Rubin, 1977, 1990) are cited as one of the most dangerous threats for causal inference in multilevel designs by a number of authors (Gitelman, 2005; Hong & Raudenbush, 2006; Sobel, 2006; VanderWeele, 2008). However, inconsistencies remain in the assessment of consequences of SUTVA violations and possible alternative assumptions. No consensus has been reached in the literature if, when, and how interactions between units within a cluster and within and between treatment groups within a cluster can be modeled in a way as to still allow for meaningful causal inferences. We will further discuss the different conceptions and show how

they can be reconciled in Chapter 3.

The second area of disagreement is the status of the cluster variable: Some authors implicitly consider the cluster variable as a function of the individual unit (Hong & Raudenbush, 2006; VanderWeele, 2008; Sobel, 2006) and confine all inferences to the current allocation of units to clusters. Others explicitly acknowledge that units could potentially be assigned to more than one cluster and that the cluster variable itself might influence the potential-outcome differentially for each unit (Gitelman, 2005), leading to the introduction of cluster-specific potential-outcome variables. While these different conceptions may be partially due to the subject matter the respective authors applied the potential-outcome framework to, correctly accounting for the status of the cluster variable remains of critical importance for a general account of causal inference for multilevel designs. While the actual status of the cluster variable is in fact determined by the research question and the population under consideration, a more general account and clarification of the the status of the cluster variable is warranted. We will take up these questions again in the definition of the single-unit trials in Chapter 2 and discuss it with respect to different designs and SUTVA violations again in Chapter 3.

Finally, attempts to identify the average causal effect with parameters in multilevel linear models or with other adjustment methods have been incomplete. They are either confined to designs with randomization at the unit- or the cluster-level (Moerbeek et al., 2000; Schochet, 2008; Raudenbush & Liu, 2000), do not refer to an explicit theory of causal effects (Seltzer, 2004) or do not explicitly or only insufficiently account for interactions between covariates and the treatment variable (Gitelman, 2005; VanderWeele, 2008). Implementations of propensity score modeling (Hong & Raudenbush, 2006) or weighting procedures (Hong & Raudenbush, 2008; Raudenbush et al., 2006) have also not detailed the problem of interactions between treatment and covariates and are unspecific on how to account for the clustered structure of the data. In general, the discussion of statistical implementations of adjustment models has not received sufficient attention in the literature. This is especially problematic in light of recent developments in the analysis of causal effects for singlelevel designs that show that the conventional general linear model might lead to biased standard errors when used as means to estimate and test average causal effects in the generalized ANCOVA (Kröhne, 2009; Nagengast, 2006) and the emergence of new statistical methods to estimate contextual effects (Croon & van Veldhoven, 2007; Lüdtke et al., 2008). We will shed further light on these questions in Chapters 4 and 5.

1.4 Outlook

In the remainder of this thesis, we will try to address the shortcomings of the literature of causal inference for between-group multilevel designs: In Chapter 2, we will introduce a general theory of causal effects (Steyer et al., 2009). We will show how specifics of multilevel designs can naturally be represented within this framework and how the theoretical concepts therein are related to different conceptions of causal effects in the literature. In Chapter 3 we will clarify which stability assumptions are made in the general theory of causal effects. We will also show how the different conditions and definitions in the literature can be reconciled within this framework and how they depend on assumptions about the specific design that is being analyzed. Based on these discussions, we will introduce a taxonomy of multilevel designs that are covered by the theory as presented here.

After having introduced a proper theoretical framework, we will use the concepts developed therein to further study the identification and analysis of causal effects in conditionally randomized and quasi-experimental multilevel designs in Chapters 4 and 5. In Chapter 4, we will develop the generalized ANCOVA (Steyer et al., 2009) for non-randomized multilevel designs with treatment assignment at the unit-level and show how the average causal effect can be identified in the presence of treatment-covariate interactions. We will then compare different statistical implementations of this model with regard to their finite sample performance in a simulation study under realistic conditions and illustrate the model with an empirical example. After that, we will turn to designs with treatment assignment at the cluster-level in Chapter 5. Once again, we will develop the generalized ANCOVA based on linear effect functions and show how the average causal effect can be identified in the presence of treatment-covariate interactions. A simulation study compares several possible statistical implementations of this model. Once again, the application of the model implementations is illustrated with an empirical example. The thesis concludes with the general discussion in Chapter 6 where we critically discuss its merits and shortcomings and outline open research questions.

2 Causal Effects – A General Theory

In this chapter, we outline the core concepts of the theory of causal effects as introduced by Steyer et al. (2009) and show how multilevel between-group designs and their intricacies are represented within this theory. The range of multilevel designs to which the theory can be applied will be further discussed in Chapter 3.

The general theory of causal effects that we are going to present in this chapter is better suited to formalize a theory of causal inference for multilevel designs than earlier theories of causal effects (Neyman, 1923/1990; Rubin, 1974, 1977, 1978) that only considered the unit variable U, the treatment variable X and the outcome variable Y in their definitions: Rubin (1974, 1977, 1978), for example, defined the *potential-outcome* of unit u in treatment condition j deterministically as the value y of the outcome variable Y that would have been observed, had the unit been assigned to treatment condition j. Neyman (1923/1990) — and later on Steyer and colleagues (Steyer, Gabler, von Davier, & Nachtigall, 2000; Steyer, Gabler, von Davier, Nachtigall, & Buhl, 2000; Steyer et al., 2002) — had already defined *true-yields* stochastically as the expected value of the outcome variable Y given unit u and treatment condition j. Both approaches, however, assume that the outcome variable Y is not influenced systematically by any other variables besides the unit variable U and the treatment variable X, i.e., that the potential-outcomes or true-yields are *always* unbiased. Hence, both theories are implicitly restricted to covariates that are deterministic functions of the unit variable U. The general theory of causal effects, as presented here, defines its elementary building blocks — the *true-outcome variables* and *true-effect variables* — as expected values of the outcome variable Y conditional on *all* confounders, whether they are deterministic functions of the unit u or not.

As already briefly discussed in Chapter 1, this extension is required to adequately represent multilevel designs: It is theoretically possible — at least in some specific multilevel designs — that units can appear in more than one cluster. In this case, the cluster variable C is not a deterministic function of the unit variable U. If, additionally,

the cluster variable C systematically influences the outcome variable Y over and above the unit variable U and the treatment variable X, the cluster variable C can confound both the true-yields as defined by Neyman (1923/1990) and the potential-outcomes as defined by Rubin (1974, 1977, 1978). The general theory of causal effects, in which the *true-outcome variables* and the *true-effect variables* are defined conditional upon all potential confounders, including the cluster variable C and cluster-covariates V, can deal with these effects naturally: The values of all potential confounders — also referred to as the *atomic strata* — represent the most fine-grained reference for definitions of causal effects that cannot be further confounded by any pre-treatment events. In order to define these concepts precisely, we will introduce two single-unit trials and a causality space that includes the confounder σ-algebra $\mathfrak{C}_X$ generated by the set of *all* potential confounders. This σ-algebra will be used to define the true-effect variable as the effect variable that considers all possible causes of Y.

While useful as building blocks of the theory, the values of the true-effect variables can never be observed in practice, since each unit can only be observed in one treatment condition at a time. This fundamental problem of causal inference (Holland, 1986) is partly circumvented by defining *average causal effects* as the expected values of the true-effects variables over the unconditional distribution of the confounders. In contrast to the true-effect variable, the average causal effect can be identified with the empirically estimable prima-facie effects under some conditions. Conditioning on the atomic strata in the definition of the true-effect variables and then taking the expected values over the distribution of the strata is referred to as the *principle of atomic stratification* by Steyer et al. (2009). Similarly, *conditional causal effects* are defined by the corresponding conditional expected values of the true-effect variables. Again, some of these conditional causal effects can be identified in applications and — depending on the research questions — provide specific information that might be of interest in evaluation studies. Furthermore, within the general theory of causal effects, special conditional causal effects similar to Rubin's (1974, 1977, 1978) potential-outcomes, Neyman's (1923/1990) true-yields, Gitelman's (2005) cluster-specific potential-outcomes can be defined that are — in contrast to these earlier concepts — unbiased by definition though not directly identifiable.

The chapter is structured as follows: We start by outlining two single-unit trials — random experiments that capture the theoretical structure of multilevel between-group experiments. We then introduce the probability space, the filtration of σ-algebras and

the random variables that together make up the *causality space* that we will be considering in the remainder of the thesis. After a short discussion of some specificities of multilevel random experiments, the central concepts for causal inference — *true-effect variables*, *average causal effects* and *conditional causal effects* — will be introduced. Next, we show how *individual causal effects* (Neyman, 1923/1990; Rubin, 1974, 1977, 1978) and *cluster-specific individual causal effects* (Gitelman, 2005) are defined. Then, we introduce the prima-facie effects, the theoretical concepts that are empirically identifiable without further assumptions. Finally, we will discuss *unbiasedness* of prima-facie effects as a prerequisite for identification of average causal effects in experimental and quasi-experimental multilevel designs and introduce three *sufficient conditions for unbiasedness* — stochastic independence, homogeneity and unconfoundedness. The chapter closes with some remarks concerning the advantages of the presented framework in comparison to other theories of causal effects.

2.1 Single-Unit Trials

In the subsequent section, we will introduce two generic single-unit trials — random experiments designed to capture the peculiarities of between-group multilevel designs. The two single-unit trials are characterized by stochastic dependencies between events and random variables and are the necessary and sufficient background for introducing the concepts of causal effects. The single-unit trials should not be confused with the statistical models used for parameter estimation and hypothesis testing that capture the laws that govern the repetitions of the single-unit trial that constitute samples. However, it is sufficient for introducing causal effects and studying conditions for their identification with empirically estimable quantities from the pre-facto perspective, i.e., before the random experiment is conducted (see also, Steyer et al., 2009).

The distinction between the single-unit trial and the sampling model can be exemplified with simple random experiment of tossing a fair coin: The probability of heads is equal to 0.5 in this single-unit trial, if the coin is in fact fair. This probability is well-defined prior to flipping the coin and even if it is never actually flipped. In order to estimate the probability of tossing head, a sample consisting of independent and identical repetitions of the single-unit trial must be obtained. The probability for head can then be estimated by the relative frequency of heads in this sample. We will re-

turn to this distinction of the single-unit trial and sampling models again in Section 3.1, and emphasize again that the single-unit trial is sufficient for the definitions of the core concepts of the theory of general causal effects.

The two single-unit trials for multilevel designs differ in the temporal order by which units and clusters are drawn. This differentiation is necessary to cover multilevel designs that use pre-existing clusters (such as neighborhoods or schools) as well as multilevel designs in which units are assigned to clusters based on a non-deterministic assignment function. We will use the single-unit trials to develop the theory of causal effects for multilevel designs following the presentation in Steyer et al. (2009). While the single-unit trials represent the two most common classes of between-group multilevel designs and capture their specifics in general, each application actually requires the careful specification of a particular single-unit trial that details and substantiates the generic classes with respect to the considered variables and the temporal order of events. Some of the designs captured within this framework will be discussed in Section 3.2.

2.1.1 Pre-Existing Clusters

The single-unit trial for multilevel designs with pre-existing clusters follows the sampling model conventionally assumed for hierarchical linear models (Snijders & Bosker, 1999). It consists of:

(a) Sampling a cluster c (e.g., a classroom) from a population of clusters,

(b) assessing the values $v_1, \ldots, v_R$ of the cluster-covariates $V_1, \ldots, V_R$, $R \geq 1$,

(c) sampling a unit u (e.g., a student) from the population of units within the cluster,

(d) assessing the values $z_1, \ldots, z_Q$ of the unit-covariate $Z_1, \ldots, Z_Q$, $Q \geq 1$,

(e) assigning (or observing the assignment) of the unit u or the cluster c to one of several treatment conditions (represented by the value j of the treatment variable X),

(f) recording the numerical value y of the outcome variable Y.

No further assumptions about the probabilities of selecting a specific cluster — the unconditional cluster probabilities $P(C{=}c)$ — or the probabilities of selecting a unit

from a specific cluster — the conditional probabilities $P(U{=}u\,|\,C{=}c)$ — are made. Nevertheless, these probabilities are part of the distribution to which all other inferences refer. The cluster-covariates $V_1, \ldots, V_R$ represent quantitative or qualitative attributes of the cluster and may be uni- or multivariate. In further discussions, we will simply use the abbreviation V to refer to the vector of cluster-covariates. The unit u actually represents the unit at the time of sampling units from clusters. By assuming that the cluster c is sampled first, and that the unit is only sampled at a later timepoint from the units within the cluster, influences of the cluster on the unit are possible. The unit sampled in step (c) is not necessarily similar to the unit that could have been sampled at the time of the selection of a cluster c. The unit variable U, that we will introduce in Section 2.2.3 has to be interpreted accordingly as representing the unit at the time of sampling units from clusters in this single-unit trial. Formally, because of the assumption that each unit u can appear only in one specific cluster, the cluster variable C is a deterministic function of the unit variable U, although the cluster c is sampled before the unit u. The unit-covariates $Z_1, \ldots, Z_Q$ include uni- and multivariate, quantitative and qualitative, manifest and latent covariates. In further discussions, we will use the abbreviation Z to refer to the vector of unit-covariates. With respect to the treatment assignment, we will later distinguish between designs with treatment assignment at the unit-level and designs where the cluster as-a-whole is assigned to a treatment condition. As we will discuss in more detail in Chapters 3 and 5, the treatment variable X is a variable at the cluster-level and only cluster-covariates are potential confounders in the second class of designs.

By explicitly including the assessments of the cluster-covariate V and the unit-covariate Z in the single-unit trial, we have introduced it in the most general way and accounted for the inclusion of fallible covariates that are not functions of the cluster variable C or the unit variable U. However, the single-unit trial simplifies considerably, if there are no time-lags between the sampling of a cluster, the sampling of a unit from the cluster and treatment assignment. In this case, the cluster variable C and the unit variable U represent units and clusters at the onset of treatment and all covariates are either functions of the cluster variable C or the unit variable U. Steps (b) and (d) of the single-unit trial can then be omitted, since the values of the cluster-covariate V and the unit-covariate Z are determined by the selected cluster c and the selected unit u.

On contrast, the cluster-covariates V are no longer necessarily deterministic functions of the cluster variable C, if there are time-lags between the initial sampling of

clusters and the assessment of the cluster-covariates. In this case, it is possible that attributes of the cluster change between the timepoints — e.g., by adding new students to a classroom or by new inhabitants moving to a neighborhood — and the initially selected cluster c is different from the cluster at the sampling of a unit or the cluster at the onset of treatment. Such effects are captured by including cluster-covariates V that are not functions of the cluster variable C, i.e., not properties of the cluster at the initial sampling of clusters. Apart from designs with time-lags between the sampling of a cluster and the assessment of the cluster-covariates, the explicit inclusion of the cluster-covariate V is also necessary if latent variables are among the covariates at the cluster-level. Most notably, the cluster-specific treatment probabilities $P(X{=}j \mid C{=}c)$ or the regression $E(Z \mid C)$ of the unit-covariates on the cluster variable are latent cluster-covariates. In applications, their values are usually not directly assessable, but have to be estimated by their respective fallible indicators, the treatment proportion per cluster or the empirical cluster means. While the true values of these variables are determined by the cluster c, their fallible measures are cluster-covariates V that are not functions of the cluster variable C.

A similar point can be made with respect to the unit-covariates Z: If there are time-lags between the sampling of a unit from the cluster and the assessment of the unit-covariate or the onset of the treatment, the unit sampled from the cluster in step (c) might differ from the unit at the assessment of the unit-covariate Z and from the unit at the onset of treatment — e.g., by being in a different mood or experiencing a stressful life-event. Such effects are again captured by introducing unit-covariates Z that are not functions of the unit variable and vary conditional on U. In contrast to earlier theories of causality (Neyman, 1923/1990; Rubin, 1974, 1977, 1978), the effects of these covariates on the outcome and on treatment assignment can be accounted for by defining causal effects conditional upon *all* potential confounders within the general theory of causal effects.

2.1.2 Assignment to Clusters

An alternative single-unit trial for multilevel designs is appropriate for designs in which units are assigned to clusters based on a non-deterministic assignment function. It consists of:

(a) Sampling a unit u (e.g., a person) from a population of units,

(b) assessing the values $z_1, \ldots, z_Q$ of the unit-covariate $Z_1, \ldots, Z_Q$, $Q \geq 1$,

(c) assigning the unit (or observing its assignment) to one of several clusters (e.g., therapy groups, represented by the value c of the cluster variable C),

(d) assessing the values $v_1, \ldots, v_R$ of the cluster-covariates $V_1, \ldots, V_R$, $R \geq 1$,

(e) assigning (or observing the assignment) of the unit u or the cluster c to one of several treatment conditions (represented by the value j of the treatment variable X),

(f) recording the numerical value y of the outcome variable Y.

No further assumption about the sampling probabilities $P(U{=}u)$ of the units are made. They are, however, part of the distribution to which all inferences will refer. In contrast to the single-unit trial for designs with pre-existing clusters introduced in the previous section, the interpretation of the unit u is now slightly different: The unit u actually represents the unit at the time of the initial sampling units from the population with all its corresponding properties. The covariates $Z_1, \ldots, Z_Q$ at the unit-level capture changes and developments of the unit in-between the initial sampling and the assignment to clusters and may be univariate or multivariate and comprise quantitative or qualitative attributes of the unit. We will refer to the vector of unit-covariates as Z in all further discussions. No additional assumptions are made about the assignment probabilities of units to clusters $P(C{=}c \mid U{=}u)$: Each unit u can theoretically appear in more than one, but not necessarily in all clusters c. Once again, it is important to note that the probabilities $P(C{=}c \mid U{=}u)$ are part of the multilevel random experiment to which all inferences refer. The cluster-covariates $V_1, \ldots, V_R$ represent attributes of the cluster and can be univariate, multivariate, quantitative and qualitative. With respect to the treatment assignment in step (e), we will later distinguish between designs with treatment assignment at the unit-level and designs where the cluster as-a-whole is assigned to a treatment condition.

By explicitly including the assessments of the unit-covariate Z and the cluster-covariate V in the single-unit trial, we have again described it in the most general way and accounted for the inclusion of fallible covariates that are not functions of the unit variable U or the cluster variable C. However, also this second single-unit trial will simplify considerably, if there are no lags between the sampling of a unit, the assignment to a

cluster and the treatment assignment. In this case, the unit variable U and the cluster variable C represent units and clusters at the onset of treatment and all covariates are either deterministic functions of the unit variable U or the cluster variable C and their values are determined by the selected unit u and the assigned cluster c.

However, if there are time-lags between the initial sampling of a unit and the assessment of the unit-covariate or the assignment of the unit to a cluster, the unit-covariates Z are no longer necessarily functions of the units at the initial sampling. In this case, it is possible that attributes of the unit change between the timepoints — e.g., by meeting a romantic partner, undergoing surgery, etc. — and the initially sampled unit u is different from the unit at the assignment to a cluster or the unit at the onset of treatment. Such effects are captured by the unit-covariates Z that are not functions of the unit variable U that represents the units and their properties at the initial sampling. A similar point can be made, once more, with respect to the cluster-covariates V: If there are time-lags between the assignment of units to clusters, the assessment of the cluster-covariate and the assignment to a treatment condition, the cluster-covariates V are no longer necessarily functions of the cluster variable C, the cluster at the assignment of the unit in step (c) might differ from the cluster at the assessment of the unit-covariate V and from the cluster at the onset of treatment — e.g., by including new members in a psychotherapy group. Again, when latent variables are considered among the covariates at the cluster-level, most notably, cluster-specific expected values of other variables (such as treatment probabilities or unit-covariates), their fallible indicators vary conditional on C and are thus included among the cluster-covariates V.

In the following section, we will more specifically discuss the formal structure of the probability space resulting from the single-unit trials, introduce random variables and discuss their properties in detail. Most of the time, the two single-unit trials will lead to the same mathematical structures; exceptions to this rule will be explicitly highlighted. In Chapter 3, we will discuss the consequences of the temporal sequence of the assignment process of units to clusters and other assumptions in more detail.

2.2 Causality Space

In the next section, we introduce the mathematical components of the causality space, that we use as a backdrop for all causal inferences in the remainder of this thesis. An

outline of the basic concepts of probability theory necessary for understanding the following discussions is given in the appendix of Steyer et al. (2009) and in Steyer (2002). General accounts of probability and measure theory are given, for example, by H. Bauer (1981) and Loève (1977, 1978).

A causality space consists of (1) a probability space $(\mathbf{\Omega}, \mathfrak{A}, P)$, (2) random variables X and Y defined upon the probability space, (3) a filtration of sub-σ-algebras of $\mathfrak{A}$ and (4) a confounder σ-algebra $\mathfrak{C}_X$. Furthermore, the putative cause X — whose effects on the outcome variable Y are to be interpreted causally — is pre-ordered to the outcome variable Y (Steyer et al., 2009). In addition to the elements of the causality space, we will also introduce other random variables defined upon the probability space, that we will use in the remainder of our thesis.

2.2.1 Probability Space

The first element of the causality space is a probability space (H. Bauer, 1981; Steyer et al., 2009) that captures the structure of the random experiment introduced informally in the single-unit-trials in the previous section. A probability space is an indispensable component of every stochastic model — no proposition about random variables, their distributions, expected values, variances or correlations can be made without explicit or implicit reference to a probability space.

A probability space consists of three distinct components (H. Bauer, 1981): (1) the set of possible outcomes $\mathbf{\Omega}$, (2) the set of potential events $\mathfrak{A}$ that is a σ-algebra on $\mathbf{\Omega}$, i.e., a set of subsets of $\mathbf{\Omega}$ that is closed with respect to unions and complements and implicitly to intersections, and (3) the probability measure P that assigns a value in the interval $[0, 1]$ to every event in $\mathfrak{A}$. This mapping must follow the axioms of probability: positivity, σ-additivity and standardization (Kolmogorov, 1933/1956). The triple $(\Omega, \mathfrak{A}, P)$ is called a probability space.

For the two single-unit trials introduced in the previous section, the set of possible outcomes, in its most general form, is the Cartesian product

$$\mathbf{\Omega} = \Omega_U \times \Omega_Z \times \Omega_C \times \Omega_V \times \Omega_X \times \Omega_Y, \tag{2.1}$$

where Ω_U is the set of observational units or the population of units, Ω_Z is the set of covariates at the unit-level that still vary given a unit u, Ω_C is the set of clusters or

the population of clusters, Ω_V is the set of covariates at the cluster-level that still vary given a cluster c, Ω_X is the set of treatment conditions and Ω_Y is the set of outcome components. For single-unit trials that only include covariates that are functions of the unit variable U and the cluster variable C, e.g., single-unit trials with no time lags between the selection of units and clusters and the onset of the treatment, the set of possible outcomes defined in Equation 2.1 is reduced by omitting Ω_Z and Ω_V:

$$\mathbf{\Omega} = \Omega_U \times \Omega_C \times \Omega_X \times \Omega_Y. \tag{2.2}$$

Since all unit-covariates are determined by the unit u and all cluster-covariates are determined by the cluster c in such designs, Ω_Z and Ω_V can be omitted in Equation (2.2). The definition of the probability space does not imply a temporal order of the events in $\mathfrak{A}$, hence, single-unit trials for designs with pre-existing clusters and for designs with assignment of units to clusters are captured by similar probability spaces.

2.2.2 Filtration

A probability space, in itself, does not imply a time order of the events. However, the two single-unit trials introduced in Section 2.1 explicitly assumed temporal sequences of events. In order to represent this time order, the concept of a filtration $(\mathfrak{C}_t)_{t\in T}$, a set of monotonically increasing sub-σ-algebras contained in the set of potential events $\mathfrak{A}$, is introduced.

Specifically, we assume that the set of potential events $\mathfrak{A}$ of the probability space $(\Omega, \mathfrak{A}, P)$ contains a filtration $(\mathfrak{C}_t)_{t\in T}$ of three σ-algebras (for further discussions, see, H. Bauer, 1981; Steyer et al., 2009). We will conceive of $\mathfrak{C}_1$ as the σ-algebra generated by all pre-treatment events, $\mathfrak{C}_2$ is the σ-algebra generated by all pre-treatment events and the treatment events and $\mathfrak{C}_3$ is the σ-algebra generated by all pre-treatment events, the treatment and the outcome events. We assume that $\mathfrak{C}_1$ is pre-ordered to both $\mathfrak{C}_2$ and $\mathfrak{C}_3$ and that $\mathfrak{C}_2$ is pre-ordered to $\mathfrak{C}_3$ (see, Steyer, 1992; Steyer et al., 2009, for a formal definition of the pre-orderedness relation). The same pre-orderedness relation holds for all random variables measurable with respect to the corresponding σ-algebras that we will introduce in the next section. Throughout this thesis, we will assume that the confounder σ-algebra $\mathfrak{C}_X$ is equal to $\mathfrak{C}_1$, the σ-algebra of all pre-treatment events (Steyer et al., 2009).

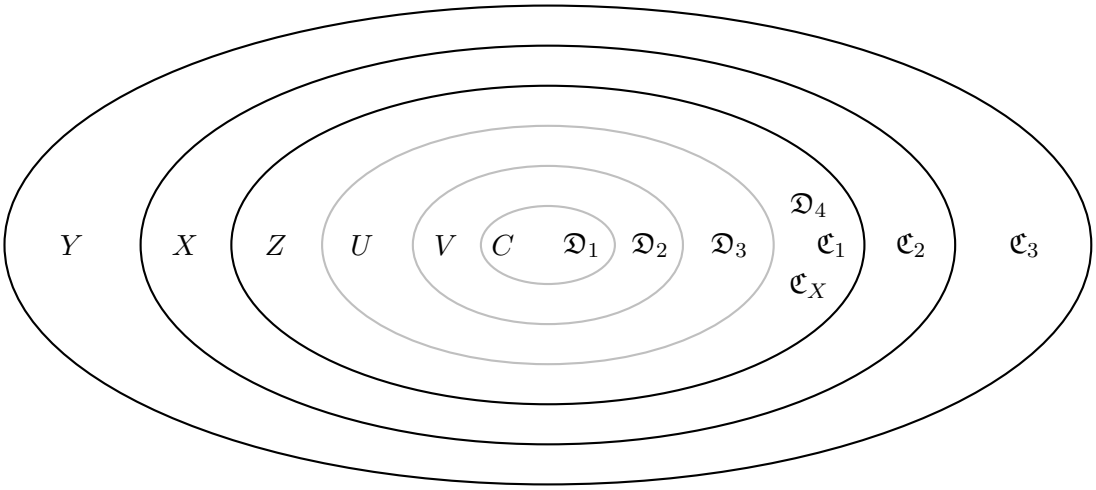

Figure 2.1: Venn-diagram of the filtrations $(\mathfrak{C}_t)_{t\in T}$ and $(\mathfrak{D}_t)_{t\in T}$ of the single-unit trial for designs with pre-existing clusters. The random variables Y, X, Z, U, V and C are presented within the circles of the σ-algebra with respect to which they are first measurable.

The two single-unit trials introduced above do not differ in the filtration considered so far. However, they lead to a different time-order of pre-treatment events in $\mathfrak{C}_1$. For the multilevel single-unit trial appropriate for designs with pre-existing cluster (as introduced in Section 2.1.1) that starts with selecting a cluster c, $\mathfrak{C}_1$ contains a filtration $(\mathfrak{D}_t)_{t\in T}$ of four σ-algebras: $\mathfrak{D}_1$ is the σ-algebra generated by the cluster variable $C : \mathbf{\Omega} \to \Omega_C$, $\mathfrak{D}_2$ is the σ-algebra generated by the union of $\mathfrak{D}_1$ and the cluster-covariate $V : \mathbf{\Omega} \to \Omega_V$, $\mathfrak{D}_3$ is the σ-algebra generated by the union of $\mathfrak{D}_2$ and the unit variable $U : \mathbf{\Omega} \to \Omega_U$ and $\mathfrak{D}_4$ is the σ-algebra generated by the union of $\mathfrak{D}_3$ and the unit-covariate $Z : \mathbf{\Omega} \to \Omega_Z$. $\mathfrak{D}_4$ is obviously identical to the confounder σ-algebra $\mathfrak{C}_X$. If only covariates that are functions of the unit variable U or the cluster variable C are considered, the filtration $(\mathfrak{D}_t)_{t\in T}$ simplifies accordingly. The filtration for this single-unit trial is represented as a Venn-diagram in Figure 2.1.

For the second class of single-unit trials appropriate for designs with assignment of units to clusters as introduced in Section 2.1.2, $\mathfrak{C}_1$ contains a different filtration $(\mathfrak{F}_t)_{t\in T}$: In this case, $\mathfrak{F}_1$ is the σ-algebra generated by the unit variable $U : \mathbf{\Omega} \to \Omega_U$, $\mathfrak{F}_2$ is the σ-algebra generated by the union of $\mathfrak{F}_1$ and the unit-covariate $Z : \mathbf{\Omega} \to \Omega_Z$, $\mathfrak{F}_3$ is the σ-algebra generated by the union of $\mathfrak{F}_2$ and the cluster variable $C : \mathbf{\Omega} \to \Omega_C$ and $\mathfrak{F}_4$

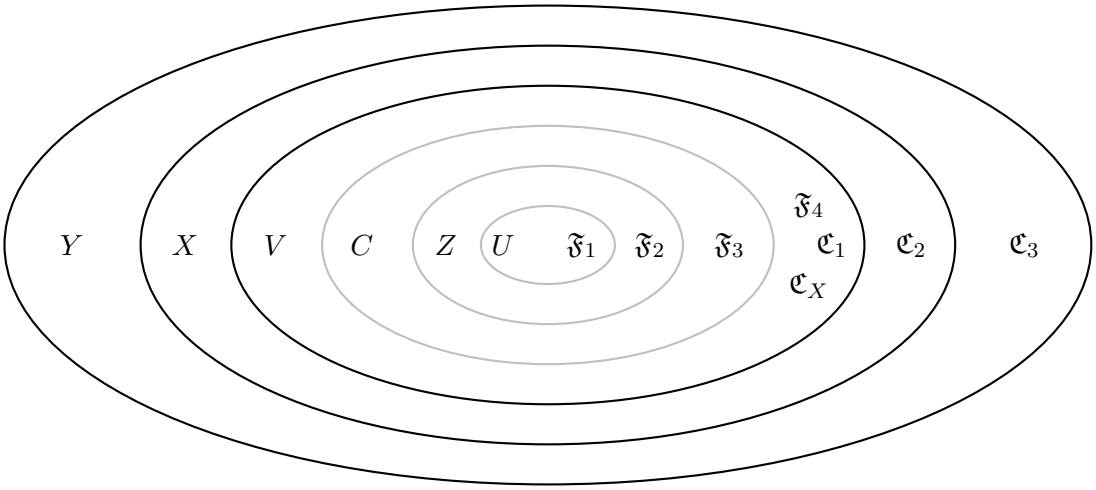

Figure 2.2: Venn-diagram of the filtrations $(\mathfrak{C}_t)_{t\in T}$ and $(\mathfrak{F}_t)_{t\in T}$ for the single-unit trial for designs with assignment of units to clusters. The random variables Y, X, V, C, Z and U are presented within the circles of the σ-algebra with respect to which they are first measurable.

is the σ-algebra generated by the union of $\mathfrak{F}_3$ and the cluster-covariate $V : \mathbf{\Omega} \to \Omega_V$. Once again, $\mathfrak{E}_F$ is identical to the confounder σ-algebra $\mathfrak{C}_X$. If only covariates that are functions of the unit variable U or the cluster variable C are considered, the filtration $(\mathfrak{F}_t)_{t\in T}$ simplifies accordingly. The filtration for this single-unit trial is represented as a Venn-diagram in Figure 2.2. While the two single-unit trials differ in the time order of pre-treatment events as captured by the two different filtrations $(\mathfrak{D}_t)_{t\in T}$ and $(\mathfrak{F}_t)_{t\in T}$, the resulting confounder σ-algebras $\mathfrak{C}_X$ are equivalent and all further derivations will apply to both single-unit trials, if not stated otherwise.

2.2.3 Random Variables

We already used random variables implicitly in the introduction of the filtrations in the previous section and will now further introduce the properties of the random variables defined upon the probability space. Most generally, random variables are defined as mappings $W : \mathbf{\Omega} \to \Omega'$, where $\mathbf{\Omega}$ is the set of possible outcomes and Ω' is a set containing as a subset the set $W(\Omega)$ of possible values of the random variable. A second condition for a random variable is the existence of a σ-algebra $\mathfrak{A}'$ whose inverse image

is a subset of $\mathfrak{A}$ (H. Bauer, 1981). In the present case, random variables will either consist of a mapping of $\mathbf{\Omega}$ onto a subset of $\mathbf{\Omega}$ or onto $\overline{\mathbb{R}}$, the set of real numbers including positive and negative infinity, to represent numerical random variables. We will start with considering the random variables that represent pre-treatment events whose generating σ-algebras are part of $\mathfrak{C}_1$ and, since $\mathfrak{C}_1 = \mathfrak{C}_X$, of the confounder σ-algebra $\mathfrak{C}_X$ and then introduce the treatment variable X and the outcome variable Y. An overview of the random variables and their properties is given in Table 2.1.

The *unit variable* U is defined as the mapping $U : \mathbf{\Omega} \rightarrow \Omega_U$, i.e., the mapping of the set of possible outcomes on the set of all units, that represents with its values which unit u is drawn. The *unit-covariate* Z is the mapping $Z : \mathbf{\Omega} \rightarrow \Omega_Z$ that represents with its values which events z have been realized. For the present purpose, we will assume that Z is a numerical random variable, i.e., that Ω_Z is equal to $\overline{\mathbb{R}}$. In some instances, e.g., when there is no time-lag between the sampling of the unit u and the assessment of the value of the unit-covariate z, the unit-covariate Z can be assumed to be a deterministic function of the unit variable U, i.e., $Z = f(U)$. However, in contrast to previous versions of the theory of causal effects (e.g., Steyer et al., 2002) this is not a necessary assumption. The *cluster variable* C is the mapping $C : \mathbf{\Omega} \rightarrow \Omega_C$ that represent with its values which cluster c has been sampled. C is assumed to be a discrete random variable, and can also be represented by M indicator variables $I_{C=c}$, where $(M + 1)$ is the total number of clusters in the population. The *cluster-covariate* V is defined as the mapping $V : \mathbf{\Omega} \rightarrow \Omega_V$ that represents with its values which events v have been realized. In line with our definition of the unit-covariate Z, we will assume that V is a numerical random variable, i.e., that Ω_V is equal to $\overline{\mathbb{R}}$. The additional assumption of V being a function of the cluster variable C, i.e., $V{=}f(C)$, is realistic in some applications, e.g., when there is no time-lag between the selection of the cluster and the assessment of the cluster-covariate, but not necessary in the further discussions. The definitions of the unit-covariate Z and the cluster-covariate V are not restricted to univariate variables, but include multivariate covariates (see Steyer et al., 2009). It should be noted, that the unit variable U, the unit-covariate Z, the cluster variable C and the cluster-covariate V are all variables measurable with respect to $\mathfrak{C}_X$. This implies that they are all potential confounders.

The *treatment variable* X is defined as the mapping $X : \mathbf{\Omega} \rightarrow \Omega_X$ that represents with its values which treatment condition j has been realized. We assume that X is a discrete numerical random variable, i.e., that Ω_X is equal to $\overline{\mathbb{R}}$ and that X can be represented with

Table 2.1: Properties of the random variables defined upon the multilevel probability space

Random variable		Mapping	Measurable[a]	Properties
U:	unit variable	$\mathbf{\Omega} \rightarrow \Omega_U$	$\mathfrak{C}_X$, $\mathfrak{C}_2$, $\mathfrak{C}_3$	
Z:	unit-covariate	$\mathbf{\Omega} \rightarrow \Omega_Z$	$\mathfrak{C}_X$, $\mathfrak{C}_2$, $\mathfrak{C}_3$	numerical
C:	cluster variable	$\mathbf{\Omega} \rightarrow \Omega_C$	$\mathfrak{C}_X$, $\mathfrak{C}_2$, $\mathfrak{C}_3$	discrete
V:	cluster-covariate	$\mathbf{\Omega} \rightarrow \Omega_V$	$\mathfrak{C}_X$, $\mathfrak{C}_2$, $\mathfrak{C}_3$	numerical
X:	treatment variable	$\mathbf{\Omega} \rightarrow \Omega_X$	$\mathfrak{C}_2$, $\mathfrak{C}_3$	discrete
Y:	outcome variable	$\mathbf{\Omega} \rightarrow \Omega_Y$	$\mathfrak{C}_3$	numerical

[a] Measurable with respect to which sub-σ-algebras of the filtration $(\mathfrak{C}_t)_{t \in T}$

J indicator variables $I_{X=j}$, where $(J+1)$ is the total number of treatment groups. Finally, the outcome variable Y is defined as the mapping $Y : \mathbf{\Omega} \rightarrow \Omega_Y$ that represents with its values which events y have been realized. We further assume that Y is a numerical random variable, i.e., that Ω_Y is equal to $\overline{\mathbb{R}}$ with finite first- and second-order moments.

Since Y is by definition measurable with respect to the σ-algebra $\mathfrak{C}_3$, but not with respect to $\mathfrak{C}_2$, the treatment variable X is pre-ordered to the outcome variable Y. The two variables thus fulfill the requirement of the causality space that the putative cause is prior to the outcome variable in the filtration. Furthermore, all random variables generated by events in $\mathfrak{C}_1$, i.e., the unit variable U, the cluster variable C, the unit-covariate Z and the cluster-covariate V are pre-ordered to both the treatment variable X and the outcome variable Y. For an extensive discussion of the concept of pre-orderedness of random variables and its relevance with respect to causal regression models see Steyer (1992) and Steyer et al. (2009).

Based upon the probability space and the definition of random variables, a large number of conditional probability and density functions can be defined that characterize the joint and conditional distributions of these random variables. We will not introduce them formally in detail here, but note that all random variables have a joint distribution. All causal inferences refer to the single-unit trial and the dependencies therein.

2.2.4 Conclusion

In the preceding section, we have introduced the mathematical components of the causality space (Steyer et al., 2009) that we will be considering in the rest of this thesis

and to which all causal inferences will refer. The causality space consists of (1) the probability space $(\boldsymbol{\Omega}, \mathfrak{A}, P)$ for the single-unit trials, (2) the filtration of sub-σ-algebras $\mathfrak{C}_1$, $\mathfrak{C}_2$ and $\mathfrak{C}_3$, (3) the treatment variable X and the outcome variable Y as defined above and (4) the confounder σ-algebra $\mathfrak{C}_X$ that is, in this case, equal to $\mathfrak{C}_1$, the σ-algebra induced by all pre-treatment events. We also already noted that the treatment variable X is pre-ordered to the outcome variable Y due to the pre-orderedness of the two σ-algebras $\mathfrak{C}_2$ and $\mathfrak{C}_3$ in the filtration. We will refer to the quintuple $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ with the specifications as described above as multilevel causality space and use it as a basis for all further definitions and derivations.

2.3 Multilevel Properties of Random Variables

In the subsequent section, we will introduce additional properties of random variables in multilevel random experiments. We start with introducing the decomposition of variables in *between-* and *within*-cluster components and briefly mention some of the ensuing consequences. In addition, we define the intraclass correlation coefficient (*ICC*) as an important descriptive concept for multilevel analyses. Finally, we discuss between- and within-cluster regressive dependencies.

2.3.1 Decomposition of Variables

The decomposition of unit-level variables into *between-cluster* and *within-cluster components* is an important property of the probability space for multilevel designs. We will discuss this property very generally referring to an arbitrary numerical random variable W, its regression on the cluster variable $E(W \mid C)$ and the corresponding within-cluster residual variable W_w. The principle applies similarly to the outcome variable Y, the unit-covariate Z, the cluster-covariate V and the treatment variable X and is useful for distinguishing between variables at the unit- and at the cluster-level.

Each numerical random variable W with a finite expected value in the random experiment introduced in Section 2.2.3 can always be decomposed into its regression on the cluster variable $E(W \mid C)$ and a within-cluster residual $W_w \equiv W - E(W \mid C)$:

$$W = E(W \mid C) + W_w = W_b + W_w, \tag{2.3}$$

where W_b is used as abbreviation for the regression $E(W \mid C)$. This decomposition is well known in statistical multilevel modeling and contextual analysis (e.g., Croon & van Veldhoven, 2007; Lüdtke et al., 2008; B. O. Muthén, 1998-2004; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Using conventional terminology (e.g., Snijders & Bosker, 1999), we will also refer to $E(W \mid C)$ as the *between-component* of W — using the expression W_b as abbreviation for $E(W \mid C)$ — and to the residual W_w as the *within-component* of W. Like every regression, $E(W \mid C)$ is a function of its regressor C, i.e., $W_b = f(C)$. The definition of the regression $E(W \mid C)$ and its residual W_w in Equation 2.3 implies that the following propositions can be derived from regression theory (cf., H. Bauer, 1981; Steyer, 2002) and hold without further assumptions. They are made explicit here for reference and use in other proofs:

$$E(W_w) = 0, \tag{2.4}$$

$$E(W_w \mid C) = 0, \tag{2.5}$$

$$Cov\left[f(C), W_w\right] = 0, \tag{2.6}$$

$$Cov\left(W_b, W_w\right) = Cov\left[E(W \mid C), W_w\right] = 0, \tag{2.7}$$

$$E(W_b) = E\big[E(W \mid C)\big] = E(W). \tag{2.8}$$

Without further assumptions, the expected value of the residual W_w is equal to zero [Equation (2.4)]. The residual W_w is regressively independent of the cluster variable C [Equation (2.5)]. This implies without further assumptions, that the covariance of every numerical function $f(C)$ of the cluster variable C and the within-component W_w is equal to zero [Equation (2.6)]. Since W_b is a function of the cluster variable C, the covariance of the between-component W_b and the within-component W_w is also equal to zero [Equation (2.7)]. Furthermore, the expected value of the between-component $E(W_b)$ is equal to the expected value $E(W)$ [Equation (2.8)]. The between-component W_b is a latent variable; the empirical cluster means $\overline{W}_{C=c}$ are fallible measures for the values of W_b whose reliabilities depend on the number of units sampled from a cluster and the intraclass correlation coefficient (Lüdtke et al., 2008). The same holds for the within-component W_w: It is only approximated in samples by the cluster-mean centered observed scores of W (Enders & Tofighi, 2007).

2.3.2 Intraclass Correlation Coefficient

An important descriptive quantity of random variables in multilevel random experiments is the *intraclass correlation coefficient* (*ICC*, Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). It can be best understood with regard to Equation 2.3. The *ICC* of any numerical random variable W with finite expected value is defined as the coefficient of determination $R^2_{W|C}$ of the regression of W on the cluster variable C:

$$ICC(W) \equiv R^2_{W|C} = \frac{Var\,[E(W \mid C)]}{Var(W)} = \frac{Var\,[E(W \mid C)]}{Var\,[E(W \mid C)] + Var(W_w)}. \tag{2.9}$$

By definition, the *ICC* of cluster-level variables that are functions of the cluster variable C is equal to one, since $Var(W_w) = 0$ for these variables. The $ICC(W)$ can also be interpreted as the reliability of W as a measure for the between-component W_b (see e.g., Bliese, 2000). The reliability of the empirical cluster means $\overline{W}_{C=c}$ as measures for the between-component W_b can be obtained by applying the Spearman-Brown formula to the reliabilities of W (Bliese, 2000; Snijders & Bosker, 1999):

$$Rel(\overline{W}_{C=c}) \equiv \frac{n_c \cdot ICC(W)}{1 + (n_c - 1) \cdot ICC(W)}, \tag{2.10}$$

where n_c is the number of units sampled from cluster $C{=}c$.

The *residual intraclass correlation coefficient* $rICC(W_1 \mid W_2)$ for an arbitrary numerical random variable W_1 with finite expected value with respect to another arbitrary variable W_2 (Snijders & Bosker, 1999) is another important concept in multilevel analysis. It is defined as

$$rICC(W_1 \mid W_2) \equiv \frac{Var\big[E(W_1 \mid W_2, C) - E(W_1 \mid W_2)\big]}{Var\big[W_1 - E(W_1 \mid W_2)\big]}. \tag{2.11}$$

The $rICC(W_1 \mid W_2)$ indicates the proportion of the variance of the residual $\varepsilon \equiv W_1 - E(W_1 \mid W_2)$ due to differences in the conditional expected values of W_1 induced by the cluster variable C.

The *design effect* is a concept closely related to the $ICC(W)$ and important for the description of multilevel data generated by repeating the single-unit trial; it is sometimes also referred to as *variance inflation factor* (*VIF*, Hox, 2002; Kish, 1965). The *VIF* captures the amount to which standard errors are underestimated in multilevel samples

if the multilevel structure is neglected. The design effect is approximately

$$VIF(W) \approx 1 + (\overline{n}_c - 1) \cdot ICC(W), \tag{2.12}$$

where $\overline{n}_c$ is the average cluster size in the sample and $ICC(W)$ is the intraclass correlation coefficient of the variable W as defined in Equation 2.9. A residual $VIF(W_1 \mid W_2)$ can also be obtained using the residual intraclass correlation coefficient $rICC(W_1 \mid W_2)$ in Equation (2.12): It approximates by which amount the standard errors are underestimated in regression models that include the covariate W_2, but ignore the cluster variable C.

2.3.3 Within- and Between-Cluster Dependencies

The decomposition of unit-level variables in between- and within-cluster components is also relevant when regressive (and stochastic) dependencies of a numerical random variable W_1 with a finite expected value and another arbitrary random variable W_2 are considered. Most basically, three types of regressions can be differentiated in a multilevel random experiment (Snijders & Bosker, 1999): (1) The *total regression* $E(W_1|W_2)$ that does not take the multilevel structure of the random experiment into account, (2) the *between-cluster regression* $E(W_{1b} \mid W_{2b})$, i.e., the regression of the between-component W_{1b} on the between-component W_{2b} and (3) the *within-cluster regression* $E(W_{1w} \mid W_{2w})$, i.e., the regression of the within-component W_{1w} on the within-component W_{2w}. The regressions at the within- and the between-cluster level can differ completely; considering only the regression at one level of aggregation can lead to erroneous conclusions, a fallacy referred to as Robinson's (1950) paradox, aggregation bias (Alker, 1969) or ecological fallacy (Greenland, 1992).

In case of linear regressive dependencies at each level, the regression coefficient β_{total} of the linear regression $E(W_1 \mid W_2)$ is a mixture of the regression coefficients $\beta_{between}$ of the linear regression $E(W_{1b} \mid W_{2b})$ and β_{within} of the linear regression $E(W_{1w} \mid W_{2w})$ (see, e.g., Snijders & Bosker, 1999, Chapter 3):

$$\beta_{total} = ICC(W_2) \cdot \beta_{between} + [1 - ICC(W_2)] \cdot \beta_{within}. \tag{2.13}$$

Considering the regression $E(W_1 \mid W_2)$ is only appropriate if the regression coefficients $\beta_{between}$ and β_{within} are equal, i.e., if the regressive dependency of W_1 on W_2 is the same

within and between clusters. Also, if the $ICC(W_2)$ is equal to zero, i.e., if W_2 is regressively independent of the cluster variable C, or if the $ICC(W_2)$ is equal to one, i.e., if W_2 is constant given the cluster variable C, it is not necessary to model the between- and within-cluster components separately (Snijders & Bosker, 1999). Otherwise, the effects of W_{2b} and W_{2w} have to be modeled separately in a multilevel regression $E(W_1 \mid W_{2b}, W_{2w})$ to avoid the risk of reaching misleading conclusions based on the regressive dependencies of W_1 on W_2. A similar decomposition applies to the regression weights of product variables that represent interaction effects (Enders & Tofighi, 2007).

To further complicate matters, models have to take into account that W_b is a latent variable whose values are only approximated (measured with error) by the empirical cluster means [see Equation (2.10)]. If the empirical cluster means are used in regression models instead of the true values of the latent variable W_b, the estimated regression weight $\hat{\beta}_{between}$ is biased (Asparouhov & Muthén, 2006; Croon & van Veldhoven, 2007; Lüdtke et al., 2008):

$$E(\hat{\beta}_{between} - \beta_{between}) = (\beta_{within} - \beta_{between}) \cdot \frac{1}{\overline{n}_c} \cdot \frac{1 - ICC(W_2)}{ICC(W_2) + [1 - ICC(W_2)]\,/\overline{n}_c}, \quad (2.14)$$

where $\overline{n}_c$ is the average sample size per cluster. The bias depends on the difference of regression weights β_{within} and $\beta_{between}$ and becomes stronger with smaller average cluster sizes and a smaller $ICC(W_2)$-values of the predictor variable W_2. Again, regression weights of product variables that use the empirical cluster means suffer from the same bias. We will show the importance of the decomposition of unit-level variables in between- and within-cluster components and the biased estimation of regressive effects of between-components for the identification and estimation of the average causal effect in Chapters 4 and 5, when we develop generalized ANCOVAs for different types of multilevel designs and and discuss their implementation in statistical models.

2.4 Causal Effects

In this section, we define several causal effects of the treatment following Steyer et al. (2009) and apply their definitions to the specifics of multilevel designs. Throughout this section, we will refer to the multilevel causality space $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ introduced in Section 2.2. We start with defining the *true-outcome variable* and the *true-effect variable* as basic building blocks of the theory of causal effects. We then in-

troduce *average causal effects* as the unconditional expected values of true-effect variables and *conditional causal effects* given a value of another variable as conditional expected values of true-effect variables. We will then show how *true-yields* and *individual causal effects* — similar to the definitions of Rubin (1974, 1977, 1978) and Neyman (1923/1990) — and *cluster-specific potential-outcomes* and *individual causal effects* — similar to the definitions of Gitelman (2005) — are represented within the theoretical framework. Finally, we specifically introduce conditional causal effects given covariates measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$.

2.4.1 True-Outcome Variables and True-Effect Variables

In the definition of both the true-outcome variable and the true-effect variable, we will refer to the σ-algebra $\mathfrak{C}_X$ with respect to which all pre-treatment variables, including the unit variable U and the cluster variable C, are measurable. We use $\mathfrak{C}_X$ to define unbiased treatment effects on the most fine-grained level — by conditioning on all potential confounders. The true-outcome variable τ_j is defined as the P-unique extension of the $\mathfrak{C}_X$-conditional expectation of the outcome variable Y given $X{=}j$ to Ω (Steyer et al., 2009):

$$\tau_j \equiv E^{\circ}_{X=j}(Y \mid \mathfrak{C}_X). \tag{2.15}$$

In contrast to the conditional expectation $E_{X=j}(Y \mid \mathfrak{C}_X)$ that is uniquely defined almost surely only on the subset of events $\{X{=}j\} \subset \mathbf{\Omega}$ and has arbitrary values for all other events, the extension is defined almost surely for *all* events in $\mathbf{\Omega}$. It is, in fact, a special version of $E_{X=j}(Y \mid \mathfrak{C}_X)$. The extension and the conditional expectation $E_{X=j}(Y \mid \mathfrak{C}_X)$ share their values within $\{X{=}j\}$. For further details and a discussion of the properties of the extension see Steyer et al. (2009).

The *true-effect variable* δ_{jk} of treatment j compared to treatment k is defined in a similar vein (see Steyer et al., 2009)

$$\delta_{jk} \equiv E^{\circ}_{X=j}(Y \mid X{=}j, \mathfrak{C}_X) - E^{\circ}_{X=k}(Y \mid X{=}k, \mathfrak{C}_X) = \tau_j - \tau_k, \tag{2.16}$$

as the difference between the true-outcome variables τ_j and τ_k in two treatment conditions. The values of the true-effect variable δ_{jk} are the most fine-grained treatment effects, because $\mathfrak{C}_X$ is the σ-algebra generated by all pre-treatment variables that could potentially confound the relation between the treatment X and the outcome variable Y.

Although, the values of the true-effect variables δ_{jk}, are the basic building blocks of the general theory of causal effects and contain the information most desired in applications (e.g., the treatment effect for a specific unit u with a specific unit-covariate z in a specific cluster c), they cannot — in general — be identified in applications, since the value y of the outcome variable Y can only be observed in one treatment condition concurrently. This problem is even exacerbated by the general definition of causal effects, since it also applies to all random variables and events measurable with respect to $\mathfrak{C}_X$. This is why, we are now turning to average causal effects and conditional causal effects that — under specific conditions — can be identified by empirically estimable quantities.

2.4.2 Average Causal Effects and Conditional Causal Effects

Using the true-effect variable δ_{jk}, we can now define the *average causal effect* ACE_{jk} of treatment j compared to treatment k as follows (Steyer et al., 2009):

$$ACE_{jk} \equiv E(\delta_{jk}). \tag{2.17}$$

The average causal effect ACE_{jk} of treatment j compared to treatment k is the expected value of the respective true-effect variable δ_{jk}. The definition of the ACE_{jk} in Equation (2.17) and the definition of δ_{jk} as the difference between the respective true-outcome variables τ_j and τ_k in Equation (2.16) imply that

$$ACE_{jk} = E(\tau_j) - E(\tau_k), \tag{2.18}$$

i.e., the average causal effect ACE_{jk} is equal to the difference of the expected values of the corresponding true-outcome variables τ_j and τ_k (cf. Steyer et al., 2009).

Conditional causal effects given the value w of an arbitrary variable W that is measurable with respect to $\mathfrak{C}_X$ are defined by Steyer et al. (2009)

$$CCE_{jk\,;\,W=w} \equiv E(\delta_{jk} \mid W{=}w), \tag{2.19}$$

as the conditional expected value of the corresponding true-effect variable δ_{jk} with respect to a value w of W.

The conditional causal effects $CCE_{jk\,;\,W=w}$ given a value w of the variable W are the

values of the *conditional causal effect function* (Steyer et al., 2009)

$$CCE_{jk;W} = E(\delta_{jk} \mid W), \tag{2.20}$$

the regression of the true-effect variable δ_{jk} on W. We will introduce conditional causal effects and conditional causal effect functions for specific variables measurable with respect to $\mathfrak{C}_X$ in the next sections.

The expected value of the conditional causal effect function is equal to the (unconditional) average causal effect ACE_{jk}, as can be easily derived using standard regression algebra (Steyer et al., 2009):

$$\begin{aligned} E(CCE_{jk;W}) &= E[E(\delta_{jk} \mid W)] && (2.21)\\ &= E(\delta_{jk}) && (2.22)\\ &= ACE_{jk}. && (2.23) \end{aligned}$$

We will take advantage of this equality in the development of adjustment models for non-randomized designs introduced in Chapters 4 and 5.

2.4.3 Individual and Cluster-Specific Causal Effects

Steyer et al. (2009) showed that *potential-outcomes*, *true-yields* and *individual causal effects* — used as building blocks in the theoretical accounts of Rubin (1974, 1977, 1978) and Neyman (1923/1990) — can be biased if there are covariates that are not functions of the unit variable. In this case the aforementioned concepts no longer constitute the most fine-grained definitions of treatment effects. Based on the true-effect variables [see Equation (2.15)], Steyer et al. introduced individual causal effects as conditional effects that are already corrected for the influence of other events and covariates. We will follow their presentation and additionally apply their logic to *cluster-specific potential-outcomes* and *cluster-specific individual causal effects* to accommodate Gitelman's (2005) definitions. At the end of the section, we will briefly discuss the consequences of the choice of the single-unit trial for the definitions and relations of these concepts.

Definitions

Neyman (1923/1990) defined the *individual causal effect* of treatment j compared to treatment k on the unit u as difference between the *true yields*, the conditional expected values of the outcome variable Y given the unit and the treatment condition (cf. Steyer et al., 2009)

$$\textit{Neyman's } ICE_{jk} \;\equiv\; E(Y \mid X{=}j, U{=}u) - E(Y \mid X{=}k, U{=}u). \tag{2.24}$$

In contrast to the true-effect variables that are defined conditional upon all potential confounders measurable with respect to $\mathfrak{C}_X$, the true yields — and, hence, the individual causal effect — as defined by Neyman (1923/1990) can be biased if there are covariates that are not functions of the unit variable (see Steyer et al., 2009, for examples). The *potential-outcome variable* introduced by Rubin (1974, 1977, 1978) is a special case of the true-yield variable $E_{X=j}(Y \mid U)$ where the conditional variances $Var_{X=j}(Y \mid U)$ are equal to zero and the outcome variable Y is deterministically determined by the unit variable U and the treatment variable X. Hence, Rubin's definition of an individual causal effect as a difference between potential outcomes suffers from the same problems as Neyman's definition.

Within the general theory of causal-effects, an unbiased *individual causal effect* $\delta_{jk\,;\,U=u}$ of treatment j compared to treatment k for unit u is defined as the $(U{=}u)$-conditional expected value of the true-effect variable δ_{jk} (Steyer et al., 2009)

$$\delta_{jk\,;\,U=u} \equiv E(\delta_{jk} \mid U{=}u) = CCE_{jk\,;\,U=u}, \tag{2.25}$$

which is equivalent to the u-conditional causal effect $CCE_{jk\,;\,U=u}$. Similarly, the *individual causal effect variable* $\delta_{jk\,;\,U}$ whose values are the individual causal effects $\delta_{jk\,;\,U=u}$ is defined as the regression of the true-effect variable δ_{jk} on the unit variable U

$$\delta_{jk\,;\,U} \equiv E(\delta_{jk} \mid U) = CCE_{jk\,;\,U}, \tag{2.26}$$

and is, conceptually, the U-conditional causal effect function $CCE_{jk\,;\,U}$.

Gitelman (2005) introduced cluster-specific potential outcomes as expected values of the outcome variable Y given the unit u, cluster c and treatment condition x. In line with Rubin's (1974, 1977, 1978) conception of potential outcome variables, she

assumed that the conditional variance $Var_{X=j}(Y \mid U, C)$ of the outcome variable Y is equal to zero and Y is deterministically determined by the unit variable U, the cluster variable C and the treatment variable X. Cluster-specific individual causal effects after Gitelman (2005) are then defined as differences between the cluster-specific potential outcomes. They capture the interaction between the cluster variable C and the unit variable U, but — since they again are defined as conditional expected values of Y only upon a subset of the potential confounders in $\mathfrak{C}_X$ — can be biased.

Again, cluster-specific individual causal effects that capture the interaction between the cluster variable C and the unit variable U and are always unbiased can be easily defined as conditional causal effects in the general theory of causality of Steyer et al. (2009): The *cluster-specific individual causal effect* $\delta_{jk\,;\,U=u,C=c}$ of treatment j compared to treatment k on unit u in cluster c is defined as the (C=c)- and (U=u)-conditional expected value of the true-effect variable δ_{jk}

$$\delta_{jk\,;\,U=u,C=c} \equiv E(\delta_{jk} \mid U{=}u, C{=}c) = CCE_{jk\,;\,U=u,\,C=c}. \tag{2.27}$$

The *cluster-specific individual causal effect variable* $\delta_{jk\,;\,U,C=c}$ is defined as the extension of the (C=c)-conditional regression of the true-effect variable δ_{jk} on the unit variable U:

$$\delta_{jk\,;\,U,C=c} \equiv E^{\circ}_{C=c}(\delta_{jk} \mid U) = CCE_{jk\,;\,U,\,C=c}. \tag{2.28}$$

Finally, the *individual causal effect cluster function* $\delta_{jk\,;\,U,C}$ is defined as the regression of the true-effect variable δ_{jk} on the unit variable U and the cluster variable C:

$$\delta_{jk\,;\,U,C} \equiv E(\delta_{jk} \mid U, C) = CCE_{jk\,;\,U,\,C}. \tag{2.29}$$

Individual causal effect functions $\delta_{jk\,;\,U,C}$ are related to their U-conditional counterparts by the following simple equation:

$$\delta_{jk\,;\,U} = E(\delta_{jk\,;\,U,C} \mid U). \tag{2.30}$$

The equality postulated in Equation (2.30) can be derived as follows:

$$\begin{aligned}
\delta_{jk\,;\,U} &= E(\delta_{jk\,;\,U,C} \mid U) && \text{Eq. (2.30)} \\
&= E\left[E\left[E^{\circ}_{X=j}(Y \mid \mathfrak{C}_X) - E^{\circ}_{X=k}(Y \mid \mathfrak{C}_X) \mid U, C\right] \mid U\right] && \text{Eq. (2.16) and (2.29)} \\
&= E\left[E^{\circ}_{X=j}(Y \mid \mathfrak{C}_X) - E^{\circ}_{X=k}(Y \mid \mathfrak{C}_X) \mid U\right] && \text{Rule (vi), Steyer (2002)} \\
&= \delta_{jk\,;\,U} && \text{Eq. (2.25).}
\end{aligned}$$

Relation to Single-Unit Trials

The choice of one of the single-unit trials introduced in Section 2.1 has consequences for the interpretation of individual causal effects $\delta_{jk\,;\,U=u}$ and cluster-specific individual causal effects $\delta_{jk\,;\,U=u,C=c}$. If the single-unit trial for pre-existing clusters (as introduced in Section 2.1.1) is considered, in which every unit can only appear within a specific cluster, the cluster-specific individual causal effects $\delta_{jk\,;\,U=u,C=c}$ are equal to the corresponding individual causal effects $\delta_{jk\,;\,U=u}$. If there are no other potential confounders that are not functions of the unit variable U — e.g., in designs with no time-lags between the sampling of the cluster, the unit and assignment to treatment conditions — the individual effect variable $\delta_{jk\,;\,U}$ is equal to the true-effect variable δ_{jk} and the unit variable U is the only potential confounder included in the confounder σ-algebra $\mathfrak{C}_X$.

If the single-unit trial for designs with assignment of units to clusters (as introduced in Section 2.1.2) is considered, the cluster-specific individual causal effects $\delta_{jk\,;\,U=u,C=c}$ can assume different values in every cluster c. If units are not assigned to clusters with equal probabilities, the cluster variable C can confound the individual causal effects $\delta_{jk\,;\,U=u}$ and has to be considered in the atomic stratification among other covariates that are not functions of the unit variable U. If there are no other potential confounders that are neither a function of the unit variable U nor of the cluster variable C — e.g., in designs in which the sampling of a unit and the assignment to clusters happens immediately before the onset of the treatment — only these two variables are potential confounders measurable with respect to $\mathfrak{C}_X$; the cluster-specific individual causal effect function $\delta_{jk\,;\,U,C}$ is equal to the true-effect variable δ_{jk}. We will return to the discussion of the consequences of the different single-unit trials for the definition of cluster-specific individual causal effects in Chapter 3.

2.4.4 Specific Conditional Effects

In this section, we will exemplify some specific conditional causal effects $CCE_{jk;W=w}$ of treatment j compared to treatment k given $W{=}w$ — originally defined in Equation (2.19) — using the random variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$. In contrast to the individual causal effects introduced in the previous section, the following conditional causal effects are empirically identifiable under some conditions. All effects will be directly defined as conditional causal effect functions similar to Equation (2.20) whose values are the corresponding conditional causal effects.

We start by introducing the *unit-covariate conditional causal effect function* $CCE_{jk;Z}$

$$CCE_{jk;Z} = E(\delta_{jk} \mid Z), \tag{2.31}$$

whose values are the average causal effects conditional on values z of the unit-covariate Z. The unit-covariate conditional causal effect function $CCE_{jk;Z}$ is defined as the regression of the true-effect variable δ_{jk} on the unit-covariate Z. The expected value of the Z-conditional causal effect function $CCE_{jk;Z}$ is equal to the average causal effect ACE_{jk}:

$$ACE_{jk} = E(CCE_{jk;Z}). \tag{2.32}$$

In the context of multilevel designs, the causal effects conditional on the cluster variable and the resulting *cluster conditional causal effect function* $CCE_{jk;C}$ are of special interest. Since C is conceptually similar to any other variable measurable with respect to $\mathfrak{C}_X$, these effects are easily defined as

$$CCE_{jk;C} \equiv E(\delta_{jk} \mid C). \tag{2.33}$$

The cluster conditional causal effect function $CCE_{jk;C}$ is simply the regression of the true-effect variable δ_{jk} on the cluster variable C. The values of this function represent the effects of treatment j compared to treatment k in different clusters. If $Var(CCE_{jk;C}) > 0$, the conditional causal effects differ between clusters; there is an interaction between the treatment variable X and the cluster variable C.

Closely related to the cluster-conditional causal effect function $CCE_{jk;C}$ and its variance is the intraclass correlation coefficient $ICC(\delta_{jk})$ of the true-effect variable δ_{jk},

which is defined in line with Equation (2.9) as

$$ICC(\delta_{jk}) = \frac{Var[E(\delta_{jk} \mid C)]}{Var(\delta_{jk})} = \frac{Var(CCE_{jk;C})}{Var(\delta_{jk})}. \tag{2.34}$$

The $ICC(\delta_{jk})$ indicates which proportion of the variance of the true-effect variable δ_{jk} is due to the regression of δ_{jk} on the cluster variable C. The higher $ICC(\delta_{jk})$, the more variance of δ_{jk} is explained by the cluster variable C, the more similar are the values of the δ_{jk} within a cluster.

While the cluster conditional causal effects may be of interest on their own in some applications — e.g., when site-specific treatment effects in multisite trials are the focus of analyses and inferences (Seltzer, 2004) — they are also important parts of adjustment procedures for multilevel designs with treatment assignment at the unit-level discussed in Chapter 4. The expected value of the cluster conditional causal effect function $CCE_{jk;C}$ is equal to the average causal effect ACE_{jk}:

$$ACE_{jk} = E(CCE_{jk;C}). \tag{2.35}$$

Three additional conditional causal effect functions need to be introduced here, because of their important role for the development of adjustment models in Chapters 4 and 5: The first is the *unit-covariate cluster conditional causal effect function* $CCE_{jk;Z,C}$:

$$CCE_{jk;Z,C} \equiv E(\delta_{jk} \mid Z, C). \tag{2.36}$$

It is defined as the regression of the true-effect variable δ_{jk} on the unit-covariate Z and the cluster variable C. The expected value of the cluster-specific conditional causal effect function $CCE_{jk;Z,C}$ is equal to the average causal effect ACE_{jk}:

$$ACE_{jk} = E(CCE_{jk;Z,C}). \tag{2.37}$$

Next, the *cluster-covariate conditional causal effect function* $CCE_{jk;V}$ is defined as

$$CCE_{jk;V} \equiv E(\delta_{jk} \mid V), \tag{2.38}$$

or as the regression of the true-effect variable δ_{jk} on the cluster-covariate V. The expected value of the cluster-covariate conditional causal effect function $CCE_{jk;V}$ is equal

to the average causal effect ACE_{jk}:

$$ACE_{jk} = E(CCE_{jk;V}). \tag{2.39}$$

Finally, we introduce the *unit-covariate cluster-covariate conditional causal effect function* $CCE_{jk;Z,V}$, which is crucial for the development of adjustment procedures in non-randomized multilevel designs, as

$$CCE_{jk;Z,V} \equiv E(\delta_{jk} \mid Z, V). \tag{2.40}$$

The unit-covariate cluster-covariate conditional causal effect function $CCE_{jk;Z,V}$ is defined as the regression of the true-effect variable on both the unit-covariate Z and the cluster-covariate V. An important descriptive concept is the residual intraclass correlation coefficient $rICC(\delta_{jk} \mid Z, V)$ of the conditional causal effect function $CCE_{jk;Z,V}$. In line with Equation (2.11), it is defined as

$$rICC(\delta_{jk} \mid Z, V) = \frac{Var\big[E(\delta_{jk} \mid Z, V, C) - E(\delta_{jk} \mid Z, V)\big]}{Var\big[\delta_{jk} - E(\delta_{jk} \mid Z, V)\big]}. \tag{2.41}$$

The residual intraclass correlation coefficient $rICC(\delta_{jk} \mid Z, V)$ indicates the proportion of the residual variance of the true-effect variable δ_{jk} that is due to the cluster variable C after taking the regressive effects of the unit-covariate Z and the cluster-covariate V into account.

Once more, the expected value of the conditional causal effect function $CCE_{jk;Z,V}$ conditional on the unit-covariate Z and the cluster-covariate V is equal to the average causal effect ACE_{jk} as defined in Equation (2.17):

$$ACE_{jk} = E(CCE_{jk;Z,V}). \tag{2.42}$$

The equalities introduced in Equations (2.32), (2.35) and (2.42) are central to the development of adjustment methods for non-randomized multilevel designs in Chapters 4 and 5. Once it is possible to identify the conditional effect function by empirically estimable quantities, the expected value of the conditional effect function and thus an identifier of the average causal effect ACE_{jk} can be obtained. This is why we introduce the prima-facie effects PFE_{jk} in the next section.

2.5 Prima-Facie Effects

While true-effect variables, average and conditional causal effects are the building blocks of the theory of causal effects and the central causal estimanda in evaluation studies, they are in general not identified by empirically estimable concepts without further assumptions. In the following section, we introduce the conditional expected values of the outcome variable Y and their differences — the *prima-facie effects* (*PFE*, Holland, 1986) — that can always be estimated by the corresponding means of the outcome variable Y in the treatment groups under the usual distributional assumptions. In Section 2.6, we will further introduce and discuss the conditions under which the *PFE*s are equal to average or conditional causal effects. Once more, we are referring to the multilevel causality space $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ introduced in Section 2.2.

2.5.1 The Unconditional Prima-Facie Effect

The prima-facie effect PFE_{jk} of treatment j compared to treatment k is defined as the difference between the conditional expected values of the outcome variable Y in treatment group j and in treatment group k (Holland, 1986; Steyer et al., 2009):

$$PFE_{jk} \equiv E(Y \mid X{=}j) - E(Y \mid X{=}k). \tag{2.43}$$

The prima-facie effect PFE_{jk} can always be estimated by the difference between the means of the outcome variable Y in the respective treatment groups. Without further assumptions, the PFE_{jk} does not have a causal interpretation; it is simply the effect of the treatment variable "at first sight" (Holland, 1986, p. 949). In fact, it can be shown that the prima-facie effect PFE_{jk} can always be decomposed into the average causal effect ACE_{jk} and two bias components, the *baseline bias*$_{jk}$ and the *effect bias*$_{jk}$, which depend on the dependencies of the true-outcome variable τ_k and the true-effect variable δ_{jk} with the treatment variable X (Steyer et al., 2009)

$$PFE_{jk} = ACE_{jk} + \textit{baseline bias}_{jk} + \textit{effect bias}_{jk}. \tag{2.44}$$

The prima-facie effect PFE_{jk} will only be equal to the average causal effect ACE_{jk}, if the *baseline bias*$_{jk}$ and the *effect bias*$_{jk}$ are both equal to zero or cancel each other out. Otherwise the PFE_{jk} is just what it name implies — an effect at first sight — but not the

average causal effect of treatment j compared to treatment k. A similar decomposition is also possible for the various conditional prima-facie effects that we will introduce next; this time into the conditional causal effects $CCE_{jk\,;\,W=w}$ and the corresponding conditional *baseline bias*$_{jk\,;\,W=w}$ and conditional *effect bias*$_{jk\,;\,W=w}$.

2.5.2 Conditional Prima-Facie Effects

In this section, we introduce a variety of conditional prima-facie effects that can be used to identify conditional causal effects under the appropriate assumptions. We start with the *prima-facie effects conditional on a value* $(Z{=}z)$ *of the unit-covariate*, $PFE_{jk\,;\,Z=z}$. They are defined as the difference between the z-conditional expected values of the outcome variable Y in treatment conditions j and k (Steyer et al., 2009):

$$PFE_{jk\,;\,Z=z} \equiv E(Y \mid X{=}j, Z{=}z) - E(Y \mid X{=}k, Z{=}z). \tag{2.45}$$

The unit-covariate conditional prima-facie effects $PFE_{jk\,;\,Z=z}$ are the values of the *unit-covariate conditional prima-facie effect function* $PFE_{jk\,;\,Z}$, that is defined as the difference between the extensions of the treatment specific regressions of Y on Z to $\boldsymbol{\Omega}$ (Steyer et al., 2009)

$$PFE_{jk\,;\,Z} \equiv E^{\circ}_{X=j}(Y \mid Z) - E^{\circ}_{X=k}(Y \mid Z). \tag{2.46}$$

As in the definition of the true-outcome variable in Equation (2.15), the extension $E^{\circ}_{X=j}(Y \mid Z)$ of the treatment specific regression $E_{X=j}(Y \mid Z)$ to $\boldsymbol{\Omega}$ is defined almost surely for all events in Ω and is, in fact, a special version of $E_{X=j}(Y \mid Z)$ to which all of the algebraic rules for conditional regressions apply.

While the definitions of unconditional and conditional prima-facie effects in Equations (2.43) and (2.45) are meaningful in the context of multilevel designs, statistical inferences about group means have to take the multilevel structure of the random experiment into account, if the intraclass correlation coefficient $ICC(Y)$ or the residual intraclass correlation coefficient $rICC(Y \mid Z)$ of the outcome variable Y are not equal to zero (see, e.g., Hedges, 2007a; Murray, 1998, 2001; Snijders & Bosker, 1999).

In order to further accommodate the clustered structure of multilevel designs, we introduce *cluster conditional prima-facie effects* $PFE_{jk\,;\,C=c}$ as the difference between the cluster specific expected values of the outcome variable Y in the treatment groups j

and k:

$$PFE_{jk\,;\,C=c} \equiv E(Y \mid X{=}j, C{=}c) - E(Y \mid X{=}k, C{=}c). \tag{2.47}$$

The cluster conditional prima-facie effects $PFE_{jk\,;\,C=c}$ can be estimated by the mean difference between treatment groups in cluster c. Similarly, the *cluster conditional prima-facie effect function* $PFE_{jk\,;\,C}$, whose values are the cluster specific prima-facie effects $PFE_{jk\,;\,C=c}$, is defined as the difference between the extensions of the treatment-group specific regressions of the outcome variable Y on the cluster variable C to Ω:

$$PFE_{jk\,;\,C} \equiv E^{\circ}_{X=j}(Y \mid C) - E^{\circ}_{X=k}(Y \mid C). \tag{2.48}$$

In a similar vein, the *cluster-covariate conditional prima-facie effects* $PFE_{jk\,;\,V=v}$ and the respective *cluster-covariate conditional prima-facie effect function* $PFE_{jk\,;\,V}$ are defined:

$$PFE_{jk\,;\,V=v} \equiv E(Y \mid X{=}j, V{=}v) - E(Y \mid X{=}k, V{=}v) \tag{2.49}$$

$$PFE_{jk\,;\,V} \equiv E^{\circ}_{X=j}(Y \mid V) - E^{\circ}_{X=k}(Y \mid V). \tag{2.50}$$

The *unit-covariate cluster conditional prima-facie effect function* $PFE_{jk\,;\,Z,C}$ is defined as:

$$PFE_{jk\,;\,Z,C} \equiv E^{\circ}_{X=j}(Y \mid Z, C) - E^{\circ}_{X=k}(Y \mid Z, C). \tag{2.51}$$

Finally, it is straightforward to include the cluster-covariate V in a definition of prima-facie effects and establish *cluster-covariate unit-covariate conditional prima-facie effect function* $PFE_{jk\,;\,Z,V}$:

$$PFE_{jk\,;\,Z,V} \equiv E^{\circ}_{X=j}(Y \mid Z, V) - E^{\circ}_{X=k}(Y \mid Z, V). \tag{2.52}$$

2.6 Unbiasedness and its Sufficient Conditions

Unbiasedness of either the treatment regression $E(Y \mid X)$ or the unit-covariate-treatment regression $E(Y \mid X, Z)$ are the weakest causality conditions under which average causal effects or conditional causal effects can be identified with the respective unconditional or conditional prima-facie effects (Steyer et al., 2009). If the treatment regression $E(Y \mid X)$ is unbiased, the difference between the means of the outcome variable Y in the

treatment and the control group is an estimator of the average causal effect. Unbiasedness of the covariate-treatment regression $E(Y \mid X, Z)$ is the precondition for applying adjustment methods such as the generalized ANCOVA and for identifying the unconditional average causal effects in conditionally randomized and quasi-experimental designs. In this section, we first define unbiasedness of both $E(Y \mid X)$ and $E(Y \mid X, Z)$. We then extend the concept of unbiasedness to the cluster-treatment regression $E(Y \mid X, C)$, the cluster-covariate-treatment regression $E(Y \mid X, V)$, the unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ and the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V)$. In the second part of the section, we will introduce and discuss three sufficient conditions for unbiasedness: (1) stochastic independence, (2) homogeneity and (3) unconfoundedness. In all definitions, we refer to the multilevel causality space $\langle(\Omega, \mathfrak{A}, P),\ (\mathfrak{C}_t)_{t\in T},\ X,\ Y,\ \mathfrak{C}_X\rangle$ introduced in Section 2.2.

2.6.1 Unbiasedness

Unbiasedness is the weakest of the causality criteria introduced by Steyer et al. (2009): It has no testable implications and does not generalize to subpopulations. Nevertheless, unbiasedness — with the additional structural components of a causality space — distinguishes a causal regression from an ordinary regression, which, in general, has no causal meaning. Under unbiasedness, the empirically estimable regressions $E(Y \mid X)$ or $E(Y \mid X, W)$ can be used to obtain average causal effects whose defining components — the true-effect variables — are not directly accessible in applications. Unbiasedness implies that the two biases introduced in Equation (2.44) cancel each other out or are both equal to zero.

Steyer et al. (2009) define *unbiasedness of the treatment regression* $E(Y \mid X)$ as follows:

$$E(Y \mid X{=}j) = E(\tau_j) \quad \text{for all values } j \text{ of } X. \tag{2.53}$$

The treatment regression $E(Y \mid X)$ is unbiased, if the $(X{=}j)$-conditional expected value of the outcome variable Y is equal to the expected value of the true-outcome variable τ_j in all treatment groups j. If the treatment regression $E(Y \mid X)$ is unbiased, the prima-facie effect PFE_{jk} — defined as the difference between two values of the treatment regression in Equation (2.43) — is also unbiased and equal to the average causal effect ACE_{jk}. This also implies that the difference in the means of the outcome variable Y in

treatment groups j and k is an unbiased estimator of the average causal effect ACE_{jk} under the usual assumption of independently and identically distributed observations.

A weaker, but important assumption, is *unbiasedness of the unit-covariate-treatment regression* $E(Y \mid X, Z)$ (Steyer et al., 2009):

$$E^{\circ}_{X=j}(Y \mid Z) = E(\tau_j \mid Z) \quad \text{a.s. for all values } j \text{ of } X. \tag{2.54}$$

The unit-covariate-treatment regression $E(Y \mid X, Z)$ is unbiased, if the extension of the $(X{=}j)$-conditional regression of the outcome variable Y on the unit-covariate Z is equal to the regression of the true-outcome variable τ_j on the unit-covariate Z in all treatment conditions j. If the unit-covariate-treatment regression $E(Y \mid X, Z)$ is unbiased, the Z-conditional prima-facie effect function $PFE_{jk;Z}$ is equal to the Z-conditional causal effect function $CCE_{jk;Z}$ and can be used to identify the values of this function.

In order to include the cluster variable C in the definitions of unbiasedness, we define *unbiasedness of the cluster-treatment regression* $E(Y \mid X, C)$ as follows:

$$E^{\circ}_{X=j}(Y \mid C) = E(\tau_j \mid C) \quad \text{for all values } j \text{ of } X. \tag{2.55}$$

The cluster-treatment regression $E(Y \mid X, C)$ is unbiased, if the extension of the $(X{=}j)$-conditional regression of the outcome variable Y on the cluster variable C is equal to the regression of the true-outcome variable τ_j on the cluster variable C for all treatment conditions j.

Another definition of unbiasedness refers to the cluster-covariate-treatment regression $E(Y \mid X, V)$. *Unbiasedness of the cluster-covariate-treatment regression* $E(Y \mid X, V)$ is defined as follows:

$$E^{\circ}_{X=j}(Y \mid V) = E(\tau_j \mid V) \quad \text{a.s. for all values } j \text{ of } X. \tag{2.56}$$

The cluster-covariate-treatment regression $E(Y \mid X, V)$ is unbiased, if the extension of the $(X{=}j)$-conditional regression of the outcome variable Y on the cluster-covariate V is equal to the regression of the true-outcome variable τ_j on the cluster-covariate V in all treatment conditions j. In this definition, we do not require unbiasedness with respect to the cluster variable C, but with respect to the cluster-covariate V. It should be noted that the between-component of the unit-covariate Z is a cluster-covariate V that is a deterministic function of the cluster variable C (see also Section 2.3).

In the next step, we introduce a definition of unbiasedness that includes the unit-covariate Z and the cluster variable C. We define *unbiasedness of the unit-covariate-cluster-treatment regression* $E(Y \mid X, Z, C)$ as:

$$E^{\circ}_{X=j}(Y \mid Z, C) = E(\tau_j \mid Z, C) \quad \text{a.s. for all values } j \text{ of } X. \tag{2.57}$$

The unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ is unbiased if the extension of the $(X{=}j)$-conditional regression of the outcome variable Y on the unit-covariate Z and the cluster variable C is equal to the regression of the true-outcome variable τ_j on the unit-covariate Z and the cluster variable C.

Finally, we introduce *unbiasedness of the unit-covariate-cluster-covariate-treatment regression* $E(Y \mid X, Z, V)$ as a slightly stronger condition for causal unbiasedness:

$$E^{\circ}_{X=j}(Y \mid Z, V) = E(\tau_j \mid Z, V) \quad \text{a.s. for all values } j \text{ of } X. \tag{2.58}$$

Unbiasedness of the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V)$ holds if the extension of the $(X{=}j)$-conditional regression of the outcome variable Y on the unit-covariate Z and the cluster-covariate V is equal to the regression of the true-outcome variable τ_j on the unit-covariate Z and the cluster-covariate V. In this final definition of unbiasedness, we do not require unbiasedness with respect to cluster variable C and the unit-covariate Z, but only with respect to the cluster-covariate V and the unit-covariate Z. Again, it should be noted that the between-component of the unit-covariate Z is a cluster-covariate V that is a deterministic function of the cluster variable C (see also Section 2.3).

We will use the definition of unbiasedness of the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V)$ extensively in the development of adjustment models for non-randomized multilevel designs in Chapters 4 and 5. In these derivations, we will use the identities of the conditional prima-facie effect functions $PFE_{jk\,;\,Z,V}$ and $PFE_{jk\,;\,Z,C}$ and the respective conditional causal effect functions $CCE_{jk\,;\,Z,V}$ and $CCE_{jk\,;\,Z,C}$ that are implied by unbiasedness of $E(Y \mid X, Z, V)$. Since the expected value of the conditional causal effect functions $CCE_{jk\,;\,W}$ is equal to the average causal effect ACE_{jk} [see Equation (2.42)] the average causal effect can be estimated, once the conditional causal effect function $CCE_{jk\,;\,W}$ has been identified with the conditional prima-facie effect function $PFE_{jk\,;\,W}$.

2.6.2 Sufficient Conditions for Unbiasedness

We will now introduce sufficient conditions for unbiasedness and discuss their relevance for causal inference in multilevel between-group designs. We will restrict our discussion to *stochastic independence* of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$, to *homogeneity* and, finally, to *unconfoundedness* as the weakest testable sufficient condition for unbiasedness. The proofs that these conditions are sufficient for unbiasedness are given in Steyer et al. (2009), as are other and weaker sufficient conditions for unbiasedness and a detailed discussion of the implicative relations between them.

Stochastic Independence

The first sufficient condition for unbiasedness of the treatment regression $E(Y \mid X)$ is *stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$*, abbreviated $X \perp \mathfrak{C}_X$ (Steyer et al., 2009)

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j) \quad \text{for all values } j \text{ of } X. \tag{2.59}$$

The discrete treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ are independent, if and only if the conditional probability functions $P(X{=}j \mid \mathfrak{C}_X)$ are constant and do not depend on $\mathfrak{C}_X$ for all treatment groups. Independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ implies unbiasedness of the treatment regression $E(Y \mid X)$ (for the proof, see Steyer et al., 2009).

A second, weaker assumption is *stochastic independence of X and the confounder σ-algebra $\mathfrak{C}_X$ conditional upon an arbitrary variable W measurable with respect to $\mathfrak{C}_X$*, abbreviated $X \perp \mathfrak{C}_X \mid W$ (Steyer et al., 2009)

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid W) \quad \text{for all values } j \text{ of } X. \tag{2.60}$$

The discrete treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ are W-conditionally independent, if and only if the conditional probability functions $P(X{=}j \mid \mathfrak{C}_X)$ only depend on W and not on other random variables or events measurable with respect to $\mathfrak{C}_X$. W-conditional independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ implies unbiasedness of the corresponding covariate-treatment regression

$E(Y \mid W, X)$. In applications, any of the variables measurable with respect to $\mathfrak{C}_X$ (or combinations thereof) can take the place of W to indicate unbiasedness of the corresponding regression. Specifically, conditional independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ with respect to the cluster variable C (abbreviated $X \perp \mathfrak{C}_X \mid C$), the unit-covariate Z and the cluster-covariate V (abbreviated $X \perp \mathfrak{C}_X \mid Z, V$) and the unit-covariate Z and the cluster variable C (abbreviated $X \perp \mathfrak{C}_X \mid Z, C$) are important conditional independence conditions that we will return to when we introduce a variety of conditionally randomized experimental designs in Chapter 3.

Substantively, both unconditional and conditional independence are important conditions for unbiasedness. Unconditional stochastic independence of X and $\mathfrak{C}_X$ holds, if units or clusters are assigned randomly and with equal treatment probabilities to treatment conditions — in this case treatment probabilities do not depend on any of the potential confounders. Conditionally randomized assignment of units or clusters to treatment conditions — assigning them to treatment conditions with equal probabilities given the values of corresponding covariates — leads to conditional stochastic independence of X and $\mathfrak{C}_X$: Treatment probabilities do not depend on any of the potential confounders conditional on the covariates used for conditional randomization. The two independence conditions are also important for quasi-experimental designs in which treatment assignment is not under the control of the experimenter: In these designs, the challenge is to identify all covariates that govern the treatment assignment process, to ensure that conditional independence holds with respect to other observed or unobserved confounders. We will discuss different randomized and conditionally randomized as well as quasi-experimental between-group multilevel designs in more detail in Chapter 3.

Homogeneity

A second sufficient condition for unbiasedness is *homogeneity of the treatment regression* $E(Y \mid X)$ *with respect to the confounder* σ*-algebra* $\mathfrak{C}_X$, abbreviated $Y \vdash \mathfrak{C}_X \mid X$ (Steyer et al., 2009)

$$E(Y \mid X, \mathfrak{C}_X) = E(Y \mid X). \tag{2.61}$$

The treatment regression $E(Y \mid X)$ is homogeneous, if the variables and events represented by the confounder σ-algebra $\mathfrak{C}_X$ do not influence the conditional expected values of the outcome variable Y given the values of the treatment variable. Homogeneity of

$E(Y \mid X)$ implies unbiasedness of $E(Y \mid X)$ (for the proof, see, Steyer et al., 2009).

Once again, homogeneity can also be defined with respect to the regression of the outcome variable Y on the treatment variable X and any other pre-treatment variable W measurable with respect to $\mathfrak{C}_X$, abbreviated $Y \vdash \mathfrak{C}_X \mid X, W$, (Steyer et al., 2009)

$$E(Y \mid X, \mathfrak{C}_X) = E(Y \mid X, W). \tag{2.62}$$

The covariate-treatment regression $E(Y|X, W)$ is W-conditionally homogeneous, if none of the random variables and events measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$ influences the conditional expected values of the outcome variable Y over and above the treatment variable X and the covariate W. Homogeneity of $E(Y \mid X, W)$ implies unbiasedness of $E(Y|X, W)$. Again, we are using the random variable W as a placeholder for any pre-treatment variable measurable with respect to $\mathfrak{C}_X$ and the corresponding pre-treatment-variable-treatment regressions. Specifically, conditional homogeneity with respect to the unit-covariate Z, the cluster-covariate V and the cluster variable C and combinations of them are important concepts in multilevel designs.

Unconfoundedness

In the two previous sections, we introduced stochastic independence and homogeneity as two sufficient conditions for unbiasedness. These conditions imply that both biases of the average causal effect — the baseline bias and the effect bias — are zero. Unconfoudedness, another sufficient condition for unbiasedness, is the weakest falsifiable condition implying unbiasedness: However, it only implies that baseline and effect bias cancel each other. In contrast to unbiasedness, unconfoundedness generalizes to subpopulations. Furthermore, unconfoundedness implies not only that the average causal effects are equivalent to the prima-facie effects, but that they are also equal to the expected values of the prima-facie effects in subpopulations — two properties that we will use in the discussion of adjustment models for designs with treatment assignment at the cluster-level in Chapter 5. Unconfoundedness and its implications are extensively discussed by Steyer et al. (2009); we will summarize the relevant parts for multilevel designs here, but refer the reader to this exhaustive treatment.

In line with Steyer et al. (2009), the regression $E(Y \mid X)$ is *unconfounded* if either

$$\begin{aligned} &(a) \qquad P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j) \quad \text{a.s.} \\ &\quad \textit{or} \\ &(b) \qquad E^{\circ}_{X=j}(Y \mid \mathfrak{C}_X) = E(Y \mid X{=}j) \quad \text{a.s.} \end{aligned} \tag{2.63}$$

holds for each value j of X. It is immediately obvious from Equation (2.63) that unconfoundedness is a weaker sufficient condition for unbiasedness than either stochastic independence of X and $\mathfrak{C}_X$ or homogeneity: Unconfoundedness holds if either one of these conditions is fulfilled in each treatment group. Consequently, stochastic independence and homogeneity imply unconfoudedness. Similarly, the conditional regression $E(Y \mid X, W)$ — where W is any arbitrary variable measurable with respect to $\mathfrak{C}_X$ — is unconfounded, if either

$$\begin{aligned} &(a) \qquad P_{W=w}(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid W{=}w) \\ &\quad \textit{or} \\ &(b) \qquad E^{\circ}_{X=j,W=w}(Y \mid \mathfrak{C}_X) = E(Y \mid X{=}j, W{=}w) \end{aligned} \tag{2.64}$$

holds for each value j of X and P_W-almost all values w of W.

An important implication of unconfoudedness of $E(Y \mid X)$ is *weak causality* (see also Steyer, 1992):

$$E(Y \mid X{=}j) = E\big[E^{\circ}_{X=j}(Y \mid W)\big] \quad \text{for all values } j \text{ of } X, \tag{2.65}$$

where W is any variable measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$. Equation (2.65) can be used to falsify unconfoundedness by testing the postulated equality for observed pre-treatment variables W. A similar proposition — also resulting in a corresponding test — is implied by unconfoudedness of $E(Y \mid X, W)$.

Average stability of the expected values $E(Y \mid X{=}j)$ and their differences, the prima-facie effects is a second important implication of unconfoudedness of the treatment regression: The expected values $E(Y \mid X{=}j)$ of the outcome variable Y in treatment condition j are always the averages of the corresponding expected values $E(Y \mid X{=}j, W{=}w)$. Therefore, the prima-facie effects PFE_{jk} are always the averages of the corresponding w-conditional prima-facie effects $PFE_{jk\,;\,W=w}$ if unconfoundedness of $E(Y \mid X)$ holds.

Analogously, unconfoundedness of the regression $E(Y \mid X, W_1)$ implies that the extension of each $(X{=}j)$-conditional regression $E_{X=j}(Y \mid W_1)$ of the outcome variable Y on the unit-covariate Z is always the regression of the corresponding extension on W_1 $E\big[E^{\circ}_{X=j}(Y \mid W_1, W_2) \mid W_1\big]$ and that the prima-facie effect function $PFE_{jk\,;\,W_1}$ is always the regression of the prima-facie effect function $PFE_{jk\,;\,W_1,W_2}$ on W_1. For our purposes, W_1 and W_2 can be any random variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$.

Finally, unconfoundedness *generalizes to subpopulations*: If $E(Y \mid X)$ is unconfounded, the regression $E(Y|X, W)$ will be also unconfounded for every potential confounder W. Similarly, if the regression $E(Y \mid X, W_1)$ is unconfounded, then, for each potential confounder W_2, the regression $E(Y \mid X, W_1, W_2)$ will be unconfounded as well. This implication can be used for testing the hypothesis of unconfoundedness. Again, for our purposes W_1 and W_2 can be any random variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$.

Other Sufficient Conditions for Unbiasedness

Further weaker conditions that imply unbiasedness of the treatment regression $E(Y \mid X)$ or unbiasedness of covariate-treatment regression $E(Y|X, W)$ include unconditional and conditional stochastic independence of the treatment variable X and the true-outcome variables τ_j — Rubin's (1974, 1977, 1978) strong ignorability condition — and unconditional and conditional regressive independence of the treatment variable X and the true-outcome variables τ_j. Both imply unbiasedness of the corresponding regressions. A detailed discussion of all sufficient conditions for unbiasedness and their interrelations is given by Steyer et al. (2009).

2.7 Conclusion

In this chapter, we have introduced the general theory for causal inferences in multilevel designs. We introduced two multilevel single-unit trials and the causality space in line with Steyer et al. (2009), discussed further properties of the distributions of random variables and events in the multilevel random experiment and adapted the definitions of average, conditional and individual causal effects to the multilevel random experiment. We then introduced unbiasedness as the weakest causality criterion under which

average and conditional effects can be identified with the empirically estimable prima-facie effects. Finally, we introduced three sufficient conditions for this identification: stochastic independence, homogeneity and unconfoundedness.

As discussed at various points throughout the chapter, the general theory of causal effects (Steyer et al., 2009) is considerably more general than the approaches by Neyman (1923/1990) and Rubin (1974, 1977, 1978) and their applications to multilevel designs (Gitelman, 2005; VanderWeele, 2008). It extends these earlier theories of causal effects in between-group designs by defining the true-effect variables δ_{jk} conditional on *all* potential confounders — no matter if they are functions of the unit-variable U or not — and does not presuppose a deterministic outcome assumption like Rubin (1974, 1977, 1978). We showed that the general theory of causality is uniquely suited to represent the peculiarities of multilevel between-group designs by the principle of atomic stratification. The earlier theories of causal effects are, in fact, special instances of the general theory (for a detailed discussion, see Steyer et al., 2009). Additionally, we showed that concepts relevant in conventional multilevel analysis (such as the intraclass correlation coefficient *ICC*) are well-defined within this theoretical framework.

The single-unit trial and the causality space do not describe a *statistical* framework for the analysis of causal effects nor do they imply a *statistical sampling model* for these endeavors. All theoretical entities in the previous chapter were defined from a pre-facto perspective and characterize the dependencies between the events and variables within the probability space, even if they are unknown and never observed in practice. Attempts to estimate them from a sample require repetitions of identical and independent single-unit trials that are stable with respect to the core parameters. Sampling models become relevant and will be considered in connection with the statistical models for the analysis of causal effects in the simulation studies in Chapters 4 and 5.

In the next chapter, we will discuss violations of the so-called stable unit treatment value assumption (Rubin, 1977, 1986, 1990) that are discussed as threat to causal inference especially relevant for multilevel designs and show how they are represented and dealt with by the general theory of causal effects. We will return to the two single-unit trials for multilevel designs and discuss their properties and consequences in more detail. We will conclude that chapter by introducing the designs which can be represented by the causality space introduced here. In Chapters 4 and 5, we will develop generalized ANCOVAs for non-randomized multilevel designs and compare different implementations of these adjustment models in statistical frameworks.

3 Causal Effects – Specifics in Multilevel Designs

In the previous chapter, we introduced the general theory of causal effects (Steyer et al., 2009) and showed how two multilevel single-unit trials that capture very generic between-group multilevel designs can be represented within this framework. In this chapter, we expand and concretize our treatment of multilevel designs. In the first part of the chapter, we discuss violations of the *Stable Unit Treatment Value Assumption* (SUTVA, Gitelman, 2005; Oakes, 2004; Rubin, 1986, 1990; Rubin et al., 2004; Sobel, 2006; VanderWeele, 2008) — as a threat to causal inference especially relevant in these designs — in relation to assumptions about the assignment process of units to clusters (Gitelman, 2005; Hong & Raudenbush, 2006, 2008). In the second part of the chapter, we introduce a taxonomy of multilevel designs and discuss the prospects of causal inference in these designs.

3.1 Stability Assumptions in Multilevel Designs

In this section, we discuss one of the key challenges to causal inference in multilevel designs: Violations of the *Stable Unit Treatment Value Assumption* (SUTVA). (Other challenges include the choice of the correct covariates in quasi-experimental designs, see Section 3.2, and the correct specification of adjustment models, see Chapters 4 and 5.) We show how the general theory of causal effects deals with the assumptions in SUTVA that are controversially discussed for multilevel designs, and how they are related to assumptions about the assignment of units to clusters made in the multilevel single-unit trials. We begin with a discussion of potential violations of SUTVA in multilevel designs (Oakes, 2004; Rubin, 1986, 1990; Rubin et al., 2004) and of the suggested alternative stability assumptions (Gitelman, 2005; Hong & Raudenbush, 2006; Sobel,

2006; VanderWeele, 2008). We then review how the general theory of causal effects introduced in the previous chapter deals with the assumptions in SUTVA. In the second part of the section, we return to the discussion of the temporal order of the unit variable U and the cluster variable C and the assignment of units to clusters (Gitelman, 2005; Hong & Raudenbush, 2006, 2008) and show how the different multilevel single-unit trials affect the stability assumptions necessary for meaningful causal inferences.

3.1.1 SUTVA Violations

The *Stable Unit Treatment Value Assumption* (SUTVA) is routinely made in accounts of causal inference though seldom explicitly discussed (see, Halloran & Struchiner, 1995; Morgan & Winship, 2007; Rosenbaum, 2007; Shadish, 2002, for notable exceptions). Rubin (1986) — referring to his potential-outcome framework — gives the most concise definition of SUTVA as "the ... assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive ..." (p. 961). According to Rubin (1990), SUTVA is violated most commonly if "(a) there are versions of each treatment varying in effectiveness or (b) there exists interference between units." (p. 282, see also, Halloran & Struchiner, 1995, for a discussion of interference effects in evaluations of vaccination against infectious diseases). In short, SUTVA assumes, in Rubin's terminology, that the potential-outcomes are only influenced by the treatment and the unit and do not change when the treatment assignment probabilities or actual treatment assignments of the focal unit or other units are changed. However, it remains unclear to which theoretical level Rubin's (1977, 1986, 1990) SUTVA definition refers: (1) The probability space of the single-unit trial and the distributions of events and random variables or (2) the sampling model, consisting of replications of the single-unit trial used in applications.

Within the theory of causal effects (Steyer et al., 2009), this distinction can be clearly made: The single-unit trial and the causality space are sufficient for defining the theoretical concepts for causal inference. In order to have a clearly defined reference for the estimation of causal effects from samples, it is necessary to assume that the parameters and distributions with a causal interpretation such as the average causal effect ACE_{jk} and the conditional causal effect functions $CCE_{jk;W}$ that characterize the single-unit trial do not change between repetitions of it, i.e., that the repeated single-unit trials are

equivalent at least with respect to the causal parameters. Most generally, this will be the case if independent repetitions of identical single-unit trials are considered.

In between-group multilevel designs both potential violations of Rubin's (1990) original SUTVA are likely: The first violation can be brought about by interactions between units, clusters and treatments (Gitelman, 2005), the second violation by interferences between units within a cluster (Gitelman, 2005; Hong & Raudenbush, 2006; Sobel, 2006; VanderWeele, 2008). We will now review the two challenges to Rubin's (1977, 1986, 1990) SUTVA definition with respect to the theory of causal effects introduced in the previous chapter.

The Cluster Variable as Potential Confounder

The presence of the cluster variable C as a potential confounder in multilevel designs can violate SUTVA in line with Rubin's (1990) first class of potential SUTVA violations: In multilevel designs, individual causal effects can vary — at least potentially and in designs in which the cluster variable C is not a function of the unit variable U — depending on which cluster a unit is assigned to. Potential-outcomes as defined by Rubin (1974, 1977, 1978) do no longer only depend on the treatment condition and the unit that receives the treatment, but also on the context — the cluster — in which the treatment is administered.

There are two ways to deal with this SUTVA violation: (1) One could either consider all treatment-cluster combinations as separate treatment conditions ($X{=}j$) or (2) explicitly include the cluster variable C in the definition of cluster-specific potential-outcomes and individual effects (Gitelman, 2005). Obviously, it is the second approach that is being followed in the theory of causal effects: The cluster variable C is treated as a potential confounder measurable with respect to the σ-algebra $\mathfrak{C}_X$ and, hence, is included among the variables conditional upon which the true-outcome variables τ_j are defined.

There are two major disadvantages to the first approach of considering each cluster-treatment combination as a separate treatment condition: (1) The definition of an average treatment effect for the whole population is not straightforward anymore; in fact, the number of treatment conditions would be the product of the number of actual treatment conditions and the number of clusters in the population. (2) The assumption of non-zero treatment probabilities for all units and all treatment conditions, that is neces-

sary in order for potential-outcomes to be defined, is harder to justify (see also Fiege, 2007), especially when designs with pre-existing cluster are considered. Thus, considering cluster-treatment combinations as factual treatments does not appear to be a suitable solution for the first class of SUTVA violations in multilevel designs.

Gitelman (2005) was the first to explicitly discuss the second approach to dealing with the effect of the cluster variable C as a potential confounder of individual treatment effects: She introduced cluster-specific potential-outcomes and individual effects as a means to relax SUTVA in multilevel designs. The general theory of causal effects as introduced in Chapter 2 includes extensions of Gitelman's concepts as special cases of conditional causal effects that are guaranteed to be unbiased [see Equations (2.27) to (2.29)] and additionally handles other confounders of treatment effects by atomic stratification on $\mathfrak{C}_X$: By introducing the true-outcome variable τ_j and the true-effect variable δ_{jk}, treatment effects can vary with all events and variables that are measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$ including the unit variable U and the cluster variable C. A meaningful definition of the average causal effect ACE_{jk} is retained as the expected value of the true-effect variable δ_{jk}. Confounding of the individual causal effects in the sense of Rubin (1974, 1977, 1978) by the cluster variable C (or other pre-treatment variables that are not functions of the unit variable U) can be handled directly within this framework (see Chapter 2 and Steyer et al., 2009, for an extended discussion) and the average causal effect is identified when (conditional) unbiasedness holds. Effects of the cluster variable C on the true-outcome variables τ_j are represented either by an intraclass correlation coefficient $ICC(\tau_j)$ or by a residual intraclass correlation coefficient $rICC(\tau_j \mid W)$ greater than zero.

Interferences Between Units

While the first class of SUTVA violations and its consequences have not been discussed widely for multilevel designs (except for Gitelman, 2005) and are taken care of elegantly within the general theory of causal effects, the second class of SUTVA violations — interferences between units — and possible remedies have been discussed by different authors (Gitelman, 2005; Hong & Raudenbush, 2006; Oakes, 2004; Sobel, 2006; VanderWeele, 2008). This class of potential SUTVA violations poses a special threat for multilevel designs in which interactions between units within a cluster are likely: In educational research, for example, the effects of a teaching program for a stu-

dent might depend on interactions with the other students in the classroom in which the treatment is administered (Rubin et al., 2004). The effectiveness of a psychotherapeutic group intervention that requires extensive discussions among group members can vary depending on the specific patients in the treatment group (e.g., Helgeson, Cohen, Schulz, & Yasko, 1999). Also, spill-over effects between treated and untreated units are likely in educational settings (e.g., Hong & Raudenbush, 2006, 2008; Seltzer, 2004) and neighborhood interventions (Sobel, 2006) when designs with treatment assignment at the unit-level are considered, and may threaten the interpretation of treatment effects. It is thus necessary, to clearly define which interactions between units within a cluster and within and between treatment groups within a cluster are admissible in multilevel designs and which are not.

Three distinct positions with respect to alternative stability assumptions in multilevel designs have been suggested, all using the potential-outcome framework (Rubin, 1974, 1977, 1978): (1) Gitelman (2005) advocated a *strong multilevel SUTVA*, requiring no interferences between units within a cluster over and above static group-level variables. (2) Independently, Sobel (2006) and VanderWeele (2008) proposed a *weak multilevel SUTVA* that allows interferences between units within the same cluster, but not between units in different clusters. (3) Hong and Raudenbush (2006, 2008) introduced a *weak multilevel SUTVA* with *contextual effects*, that allows potential-outcomes to be influenced by the treatment assignment probabilities of other units within a cluster as long as these influences are represented by a contextual variable, i.e., a function of the treatment assignment (probabilities) of other units within the cluster. We will briefly summarize the different positions and show how they fit in the general theory of causal effects introduced in Chapter 2.

Strong Multilevel SUTVA. Gitelman (2005) introduced the *strong multilevel SUTVA* for designs with treatment assignment at the cluster-level as "*group membership invariance assumption* ... [that] encodes the invariance of subject-level potential outcomes to the effects of other subjects" (Gitelman, 2005, p. 404, emphasis original), implying that "a subject's potential outcome may vary with different static group-level characteristics ... but not with different compositions of the group" (p. 405). Interferences between units within a cluster are assumed to be absent. Gitelman referred to such interference effects as "group-dynamic effects" (p. 401) — effects of interactions between subjects within a cluster over and above the influence of "static group-level informa-

tion ... and individual covariate information" (p. 401). According to Gitelman, the presence of group-dynamic effects threatens causal inference in designs with treatment assignment at the cluster-level and these effects are implicitly assumed to be absent when hierarchical linear models are used to estimate average causal effects.

Weak Multilevel SUTVA. Sobel (2006) and VanderWeele (2008) independently introduced a *weak multilevel SUTVA* for designs with treatment assignment at the cluster-level. VanderWeele introduced the *Neighborhood-Level Stable Unit Treatment Value Assumption* (NL-SUTVA) in his discussion of stability assumptions in neighborhood level research. NL-SUTVA "requires that an individual's outcome does not depend on the treatment assigned to other neighborhoods other than the individual's own neighborhood" (p. 1937). He specified this assumption further by noting that it does not require that "there be no treatment interaction between two individuals in the same neighborhood ... [but] rather ... that there be no treatment interaction between individuals in different neighborhoods" (p. 1938). In a similar way, the *partial interference assumption* introduced by Sobel (2006, p. 1405) assumes that there are no interferences between units in different clusters, but allows interferences between units within the same cluster.

Weak Multilevel SUTVA with Contextual Effects. Finally, Hong and Raudenbush (2006) discussed SUTVA violations in their case study on the causal effects of Kindergarten retention and introduced a variant of the *weak multilevel SUTVA* that explicitly included *contextual effects* for designs with treatment assignment at the unit-level. Similar to Sobel (2006) and VanderWeele (2008), Hong and Raudenbush assumed no interferences between units in different schools, but allowed interferences between units within schools. More importantly, they further relaxed the second component of Rubin's (1986, 1990) original SUTVA formulation for multilevel designs with treatment assignment at the unit-level: Hong and Raudenbush (2006) explicitly modeled the influence of the average treatment assignment probability in a cluster — a function of the unit-specific treatment assignment probabilities — for all units within a cluster on the potential-outcome of the focal subject (see also, Sobel, 2006, for a discussion of violations of the stronger assumption of no spill-over effects between treated and untreated units and a discussion of instrumental variable techniques to quantify this bias). Specifically, they classified each school in their sample as either a high- or a low-retention

school depending on the retention (i.e., treatment) probabilities of the students within these schools. In terms of the random experiment introduced in Chapter 2, they treated retention rate in a school as a (fallible) measure of the cluster-conditional treatment probabilities $P(X{=}j \mid C)$ and thus implicitly as a cluster-covariate V.

Summarizing the alternative stability assumptions in multilevel designs, different conceptions of admissible interactions between units within a cluster are apparent: While Gitelman (2005) conjectures that residual interferences between units within a cluster render causal inferences meaningless, such interactions are specifically allowed in the assumptions made by Hong and Raudenbush (2006), Sobel (2006) and VanderWeele (2008). These authors only exclude interferences between units in different clusters. Hong and Raudenbush go one step further and explicitly model the potential effects of such interferences and spill-over effects of units within a cluster as operating through a function of the treatment assignments at the unit-level. We will now review how these stability assumptions fare in light of the general theory of causal effects introduced in Chapter 2 with respect to the causality space $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ and with respect to repetitions of the single-unit trial to estimate causal effects in samples.

In the previous chapter, we defined the true-outcome variable τ_j [Equation (2.15)] and the true-effect variable δ_{jk} [Equation (2.16)] by stratifying on the confounder σ-algebra $\mathfrak{C}_X$ that includes all pre-treatment events and all random variables that represent these events. Among these random variables, we explicitly considered the unit-variable U, the cluster variable C, the unit-covariate Z and the cluster-covariate V. By definition, the set of random variables that is measurable with respect to $\mathfrak{C}_X$ also includes all random variables that are functions of the aforementioned random variables. Most importantly for the present discussion, functions of the cluster variable C, such as the regression $E(Z \mid C)$ of the unit-covariate Z on the cluster-variable C or cluster-specific treatment propensity functions $P(X{=}j \mid C)$ are measurable with respect to $\mathfrak{C}_X$. Hence, they are included in the atomic strata upon which the true-outcome variables τ_j and the true-effect variables δ_{jk} are defined. This implies that the true-effect variables δ_{jk} already take different compositions of the cluster into account, as long as these compositional or contextual effects with respect to the unit-covariate Z strictly operate through functions of the cluster-variable C (or other variables measurable with respect to $\mathfrak{C}_X$), e.g., if the average intelligence level in a classroom affects the educational outcomes of students over and above the individual intelligence levels by providing an enriched learning environment. It also implies that different treatment regimes in different clus-

ters — in designs with treatment assignment at the unit-level — can influence the true-effect variables δ_{jk} without invalidating the definition of causal effects, as long as these treatment regimes can be summarized as a function of the cluster variable C, e.g., when they are represented by cluster-specific treatment propensity functions. The average causal effect is always defined as the expected value of the true-effect variable δ_{jk} [see Equation (2.17)] and thus averages over potentially confounding contextual effects measurable with respect to $\mathfrak{C}_X$.

While the preceding discussions referred to the *definition* of causal effects in the multilevel single-unit trial, *estimating* the average causal effect from samples requires that the causally relevant distributions and regressions of the single-unit trial do not change if it is repeated (see also, Rosenbaum, 2007). The resulting sequential random experiment for the sampling model must consist of independent repetitions of single-unit trials that are equivalent in their causal parameters. The statistical model used to estimate causal effects has to adequately represent the structure of the repeated single-unit trials to obtain unbiased and efficient estimates of the average causal effect and their standard errors. In its strongest version, the sequence of causally stable single-unit trials consists of independent repetitions of identical single-unit trials. However, several weaker assumptions are possible and plausible with respect to the independence and stability of the single-unit trial: Dependencies and instabilities of the repeated single-unit trials are admissible, if the causal parameters of interest — average and conditional causal effects, depending on the goal of the analysis — are not affected.

Again, dependencies and instabilities with respect to focal parameters in repetitions of the single-unit trial can be illustrated with the simple random experiment of a coin toss. The focal parameter is the probability of tossing head for a specific (and fair) coin. This probability is to be estimated by repeating the coin toss. If the coin is replaced by an unfair coin in between repetitions of the coin toss and unbeknownst to the experimenter, the identity assumption is violated and the relative frequency of head will not reflect the probability of tossing head for the original coin. If this change is undertaken depending on the outcome of specific trials, the assumption of independence is violated. In both cases, no inferences about the original single-unit trial of tossing a fair coin are possible. However, if the original fair coin is replaced with a different fair coin, the focal parameter, the probability of tossing head, is not changed and is estimated without bias by the corresponding relative frequency. Returning to the discussion of multilevel designs, an example of inadmissible dependencies and changes between repetitions of

the single-unit trial are systematic changes of the conditional causal effect $CCE_{jk;C=c}$ in cluster c by the presence or absence of a specific individual within the cluster c (Gitelman, 2005) or by the actual assignment of a specific unit to either the treatment or the control condition (Sobel, 2006). Such effects are not measurable with respect to the original confounder σ-algebra $\mathfrak{C}_X$ and cannot be adequately represented by a covariate in an adjustment model.

Re-evaluating the alternative stability assumptions in light of the theory of causal effects, Gitelman's (2005) *strong multilevel SUTVA* is unnecessarily rigid. It is not necessary that potential-outcomes of different units within a cluster are independent given all unit-covariates and static properties of the cluster. As long as such interferences between units can be modeled as contextual effects, i.e., as functions of the cluster-specific distribution of unit-covariates Z, they do not invalidate the definition of causal effects in multilevel designs. This is in line with Hong and Raudenbush's (2006) approach of modeling treatment proportions in clusters as a contextual variable. Differences in treatment proportions between clusters are a random variable measurable with respect to $\mathfrak{C}_X$. The present analysis adds to Hong and Raudenbush's analysis by explicitly including elements of cluster-specific distributions of the unit-covariate Z — e.g., the expected values of Z in each cluster, the values of the between-component Z_b — and by showing that these variables are conceptually similar to cluster-covariates V. Including them in a statistical model, however, and obtaining an unbiased estimator of the average causal effect might pose special challenges (Lüdtke et al., 2008, see also Chapter 5).

While modeling cluster-conditional functions of the distribution of unit-covariates Z and the treatment variable X as contextual effects is justified by the theory of causal effects (Steyer et al., 2009), additional interferences between units in the same cluster, that are allowed by the weak multilevel SUTVA (Hong & Raudenbush, 2006, 2008; Sobel, 2006; VanderWeele, 2008) seem to threaten the validity of causal inferences, if they do not operate through a function of the cluster-specific distribution of unit-covariates or treatment propensities. To answer the question if and when such interactions may be allowed, we return to the assignment process of units to clusters and the consequences of the temporal order of the unit-variable U and the cluster variable C in the single-unit trials introduced in Section 2.1.

3.1.2 Assignment of Units to Clusters

In this section, we further study the consequences of assumptions about the assignment process of units to clusters and the choice of the multilevel single-unit trial with respect to SUTVA violations and necessary stability assumptions in multilevel designs. In Section 2.1, we introduced two single-unit trials for multilevel designs that imply different temporal orders of the unit-variable U and the cluster-variable C. Since all inferences about causal effects refer to the underlying causality space, a further discussion of these two single-unit trials and their implications is warranted.

The single-unit trial for multilevel designs with *pre-existing clusters* started with selecting a cluster c from the set of all clusters, followed by the selection of a unit u from this cluster. We explicitly assumed that each unit u could only be present in one cluster c. The single-unit trial for multilevel designs with *assignment of units to clusters* started with selecting a unit u from the set of all units, followed by assigning (or observing the assignment) of the unit to a cluster c. This distinction sets the stage for our discussion of different assignment processes of units to clusters. Specifically, we will consider three idealized assignment mechanisms: (1) *Pre-existing clusters*, in which the cluster variable C is a function of the unit variable U and no assignment in the actual sense of the word takes place, (2) *random assignment* of units to clusters, and (3) *non-random assignment* of units to clusters (potentially conditionally independent on values of the unit-covariate Z, see also, Roberts & Roberts, 2005, for a similar distinction). Conceptually, the assignment mechanism for the two latter cases is captured in the unit-specific cluster probabilities $P(C{=}c \mid U{=}u)$. As such, it is part of the probability measure that characterizes the underlying single-unit trial. We will discuss the implications of each assumption about the assignment process with regard to (1) the consequences for the definition of true-effect variables δ_{jk} and (2) the relevance of contextual effects and SUTVA violations. In doing so, we will discuss only the idealized versions of the different assignment mechanisms, but note in advance that our comments are also relevant for more realistic cases in applications.

Pre-Existing Clusters

Multilevel designs that use pre-existing clusters such as neighborhoods, hospitals, practices or pre-existing therapy groups, are represented by the single-unit trial that assumes that the cluster variable C is temporally pre-ordered to the unit-variable U (as intro-

duced in Section 2.1.1). The notion of no actual assignment process of units to clusters — all units u appear with a probability of one in a cluster c — is formally equivalent to conceiving of the cluster variable C as a deterministic function of the unit variable U

$$C = f(U). \tag{3.1}$$

In this case, there is no self- or other-selection of units to clusters to be represented in the underlying multilevel random experiment and clusters can be thought of as pre-existing, intact entities. This has direct consequences for the interpretation of the unit variable U: Since every unit can only appear in one cluster, the units under consideration are more precisely referred to as the units-within-a-cluster. For this single-unit trial, the individual causal effect cluster functions $\delta_{jk;U,C}$ [as defined in Equation (2.29)] are uniquely defined only for the actual unit-cluster combinations and thus equal to the individual causal effect variable $\delta_{jk;U}$ [as defined in Equation (2.26)]. Causal inferences are automatically restricted to the current allocation of units to clusters (Hong & Raudenbush, 2006; VanderWeele, 2008).

However, since there is only one possible allocation of units to clusters, all interference effects between units within a cluster are functions of the cluster variable C and thus measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$. For the estimation of average causal effects from samples, this has the fortunate consequence that all interactions and interferences among units within a cluster and their effects on the outcome variable are taken care of by modeling the cluster variable C or the appropriate cluster-covariate V. This echoes the assumptions made by Hong and Raudenbush (2006, 2008), Sobel (2006) and VanderWeele (2008) who allowed interference effects between units within a cluster, as long as there were no interferences between units in different clusters. The *weak multilevel SUTVA* for designs with treatment assignment at the cluster-level is appropriate for designs in which the cluster variable C is pre-ordered to the unit variable U and can be considered a function of the unit variable U, i.e., for designs that use pre-existing clusters, such as schools, classrooms, neighborhoods or hospitals and assign whole clusters to treatment conditions. If designs with treatment assignment at the unit-level are considered and the *weak multievel SUTVA* with *contextual effects* (Hong & Raudenbush, 2006) applies, i.e., treatment assignments influence the true-outcome variables τ_j through a variable measurable with respect to $\mathfrak{C}_X$ (e.g., the treatment proportion in a cluster), the definition of causal effects is not invalidated.

However, additional interference and spill-over effects between treated and untreated units are not admissible (Halloran & Struchiner, 1995; Sobel, 2006) and different approaches to quantify them may have to be considered in their presence (Manski, 1995; Rosenbaum, 2007; Sobel, 2006). In applications, when the average causal effect is inferred from repetitions of the single-unit trial, we have to assume, in any case, that the repeated single-unit trials are equivalent with respect to their causal parameters and that the statistical model correctly reflects the structure of the repeated single-unit trials. Otherwise, causal inferences lack a clearly defined basis and become meaningless.

Random Assignment

We are now turning to the second single-unit trial introduced in Section 2.1.2 and consider designs with random assignment of units to clusters. When units are assigned randomly to clusters, the probability of being assigned to a specific cluster c is equal for all units u. This implies that the probability of being assigned to a cluster c is constant and equal to the unconditional probability of this cluster:

$$P(C{=}c \mid U{=}u) = P(C{=}c) \quad \text{for all values } u \text{ of } U \text{ and } c \text{ of } C. \tag{3.2}$$

Random assignment of units to clusters has a number of desirable consequences: As is evident from Equation (3.2), it implies stochastic independence of the unit variable U and the cluster variable C. It also implies that the values of the individual causal effect cluster function $\delta_{jk;U,C}$ are defined for all unit, cluster and treatment combinations, if the additional assumptions of non-zero treatment probabilities for every unit-cluster combination

$$0 < P(X{=}j \mid U{=}u, C{=}c) < 1 \quad \text{for all values } u \text{ of } U,\ c \text{ of } C \text{ and } j \text{ of } X, \tag{3.3}$$

is fulfilled. Random assignment of units to clusters further implies that the cluster variable C is stochastically independent of all unit-covariates Z that are functions of the units at the time of assignment to clusters, which in turn directly implies regressive independence of these unit-covariates Z from the cluster variable C. Consequently, clusters are — on average — equally composed of units with different characteristics, rendering contextual effects based on the composition of clusters irrelevant. In designs with treatment assignment at the unit-level, however, treatment proportions per clus-

ter may still influence the true-outcome variables τ_j and may have to be modeled as a cluster-covariate V. Further spill-over effects between treated and untreated units are not permissible (Sobel, 2006). Also, time-lags between the assignment of units to clusters and the onset of treatment, as reflected by the presence of unit-covariates Z that are not functions of the unit variable U at the assignment to clusters, can change the composition of the cluster with respect to these covariates.

Random assignment of units to clusters does not preclude the emergence of interference effects between units within a cluster in repeated single-unit trials: If, in such repetitions, the true-outcome variables τ_j vary in a systematic way depending on the presence or absence of specific individual units within a cluster and these variations are not captured by a random variable measurable with respect to $\mathfrak{C}_X$ in the original single-unit trial (e.g., if the assumption of independence of the causal parameters in the repeated single-unit trials is violated) such interference effects remain unmodeled and can bias the estimation of the average causal effect. As discussed in the previous section, in estimating the average causal effect from samples, we are always making the implicit assumption that the observations are obtained by repeating single-unit trials that are equivalent with respect to the causally relevant distributions and parameters. If interference effects on the true-outcomes τ_j are completely captured by the cluster variable C, e.g., by making them more alike in each cluster, and this effect remains constant over replications of the single-unit trial, the validity of the effect definition is not threatened. However, such an effect would have to be included in the statistical model used to estimate the average causal effect or other causal parameters from a sample in order to obtain unbiased and efficient estimates and standard errors. If the causally-relevant parameters of the single-unit trial change between repetitions and this change cannot be modeled as an event or random variable measurable with respect to the $\mathfrak{C}_X$ — e.g., because the cluster-specific conditional effects are changed in a systematic way by the presence or absence of a specific individual within a cluster — there is no uniquely defined basis for causal inferences. Average causal effects estimated from an actual study with random assignment of units to clusters generalize to other assignments of units to clusters obtained from the same randomization scheme — provided there are no unmeasured interferences between specific units assigned to a cluster and between and within the actual treatment and control groups in designs with treatment assignment at the unit-level.

Unfortunately, random assignment of units to clusters is unlikely to ever hold in

practical applications. One can only hope to achieve this condition in studies where the assignment is completely under the control of the experimenter. Even then, random assignment might not be feasible due to practical constraints. Evaluation studies of group psychotherapy (e.g., Baldwin et al., 2005) are one example, where it seems feasible to first randomize patients to groups and then assign these groups to treatment conditions. Another example are studies, in which units are first assigned randomly to treatment conditions, but the treatment itself is administered in groups (D. J. Bauer et al., 2008). In most real world contexts however, the allocation of units to clusters is not random and we cannot assume many of the advantageous properties implied by independence of U and C to hold by fiat.

Non-Random Assignment

We are now turning to designs with non-random assignment of units to clusters. In this case, the unit variable U is still temporally pre-ordered to the cluster variable C. However, in the most general definition of non-random assignment of units to clusters, we are only assuming that the probability of being assigned to each cluster is between zero and one for every unit, but can take on any value in this interval to represent self- or other-selection of units into clusters

$$0 < P(C{=}c \mid U{=}u) < 1 \quad \text{for all values } u \text{ of } U \text{ and } c \text{ of } C. \tag{3.4}$$

This is of course an idealized conception of the assignment process: Designs in which the probabilities of being assigned to some clusters are equal to zero, but there is still a non-deterministic assignment of units to the remaining clusters with assignment probabilities smaller than one, are probably the norm rather than the exception. Think, for example, of multisite evaluation studies in which a specific treatment regime is to be evaluated in a number of hospitals: If some of these hospitals cater to overlapping patient populations, these patients have a non-zero probability of turning to one of these hospitals, while at the same time having zero probability of receiving treatment in hospitals farer away from their neighborhood (VanderWeele, 2008). However, for the sake of clarity, we are confining our discussion to the idealized cases and note that the following discussions similarly apply to more realistic designs in applications.

The identification of average causal effects becomes more difficult with non-random assignment of units to clusters: In comparison to designs with random assignment of

units to clusters, the cluster variable C becomes a potential confounder, e.g., because units may self-select to clusters according to the expected effects of the treatment in different contexts. The identification of average causal effects thus requires conditional unbiasedness assumptions that include the cluster variable C or at least the cluster-covariate V as introduced in Section 2.6 (see also, Gitelman, 2005). If one of these conditional unbiasedness conditions holds, the average causal effect can be identified as long as all contextual effects resulting from the different compositions of clusters or different treatment assignment schemes are captured by cluster-covariates V. Interference effects between units within a cluster that are not captured in variables measurable with respect to $\mathfrak{C}_X$ are excluded, as are spill-over and interference effects between treated and untreated units in designs with treatment assignment at the unit-level. Again, one has to additionally assume that the probability space is not changed with respect to its causally relevant parameters, when the single-unit trial is repeated to obtain a sample. If there are changes to the distributions and regressions of the relevant variables, causal (and all other) inferences do no longer have a meaningful basis. Inferences about average causal effects from multilevel designs with non-random assignment of units to clusters generalize over the current allocation of units to clusters to other allocations that follow from the distributions of the same multilevel random experiment as captured in the single-unit trial.

In comparison to random assignment, non-random assignment of units to clusters is more likely to hold in practical applications. It holds in all designs that do not rely on pre-existing clusters, but in which subjects can self-select into different clusters and then receive a treatment that is either assigned to individual units or to the complete cluster.

3.1.3 Conclusion

In the preceding section, we outlined alternative conceptions of SUTVA in multilevel designs and discussed three idealized assignment mechanisms of units to clusters with regard to the consequences for causal inference. We clarified that all contextual effects that are captured by random variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$ do not invalidate the definition of causal effects in multilevel designs. The identification of the average causal effect is possible as long as conditional unbiasedness holds with respect to these contextual variables (and, of course, other relevant

covariates) and the causally relevant parameters of the single-unit trial are constant over repetitions. Additional interference effects between units within a cluster that are not captured by cluster-covariates V, are only permissible in designs using pre-existing clusters. In this case, only one allocation of units to clusters is considered and interference effects are always captured by the cluster variable C. However, if more than one treatment condition is implemented in a cluster, it is still necessary to assume that there are no further spill-over effects between treated and untreated units that are not captured by cluster-covariates measurable with respect to $\mathfrak{C}_X$. Again, inferences about the average causal effect are restricted to the random experiment under consideration and therefore to the specific allocation of units to clusters. When a subset of units has non-zero probabilities of being assigned to more than one cluster, the presence of *unmodeled* interference effects between the units actually assigned to a cluster threatens the validity of causal inferences from repeated single-unit trials. As long as such interference effects operate through aspects of the cluster-specific distribution of unit-covariates or treatment propensities, e.g., their cluster-specific expected values, variances or proportions, they can be treated as covariates on the cluster level. Again, residual inferences between treated and untreated units are not permissible. If conditional unbiasedness holds with regard to these covariates, the conditional causal effect functions can be identified with the corresponding conditional prima-facie effect functions and the average causal effect can be identified as the expected value of the conditional *PFE*-function (see also Chapters 4 and 5). In the next section, we will develop a taxonomy of multilevel designs and discuss the prospects for causal inference for each of the design types in more detail.

3.2 Taxonomy of Multilevel Designs

Building upon the introduction of multilevel designs in Chapter 1, the theory of causal effects in Chapter 2 and the stability assumptions for causal inference in multilevel designs discussed in the previous section, we will now systematically introduce and discuss a variety of multilevel between-group designs. We will extend existing taxonomies of multilevel designs (e.g., Plewis & Hurry, 1998; Ukoumunne et al., 1999) by differentiating three dimensions: (1) The *assignment of units to clusters*, (2) the *level of treatment assignment* and (3) the *treatment assignment process*. More fine-

Table 3.1: Taxonomy of multilevel designs: Dimensions and levels

(1) Level of treatment assignment
(a) Assignment at the unit-level
(b) Assignment at the cluster-level
(2) Assignment of units to clusters
(a) Pre-existing clusters
(b) Randomized assignment
(c) Non-randomized (self-)assignment
(3) Treatment assignment mechanism
(a) Experimental designs
(aa) Unconditional randomized assignment
(ab) Randomized assignment conditional on covariates
Conditional on unit-covariates Z
Conditional on the cluster variable C
Conditional on cluster-covariates V
Conditional on unit-covariates Z and the cluster variable C
Conditional on unit-covariates Z and cluster-covariates V
(b) Quasi-experimental designs
(ba) Self-assignment
(bb) Other-assignment

grained differentiations with respect to the nature of the covariates (e.g., pre-tests, see Murray, 1998) are foregone. The dimensions and levels of the taxonomy are displayed in Table 3.1. Contrary to more comprehensive taxonomies of experimental and quasi-experimental singlelevel designs (e.g., Shadish et al., 2002), we will restrict our discussion to between-group designs, i.e., to designs that compare a treatment condition to a control condition with different units in each condition. Thereby, we will explicitly exclude within-units designs such as interrupted time-series or other longitudinal designs without an external control group.

We will now introduce and discuss each dimension of the taxonomy separately. We will focus especially on the treatment assignment mechanism, as this dimension has received the least attention so far. The chapter concludes with a brief evaluation of the prospects and limitations of causal inference for multilevel designs.

3.2.1 Level of Treatment Assignment

In keeping with the traditional distinction for multilevel designs (Moerbeek et al., 2000; Plewis & Hurry, 1998; Seltzer, 2004) and our discussion in Chapter 1, we will distinguish between designs with (1) *treatment assignment at the unit-level* and (2) designs with *treatment assignment at the cluster-level*.

In designs with treatment assignment at the unit-level, individual units are assigned to treatment conditions and units receive an individualized treatment (e.g., surgical procedure, individual tutoring, see also Raudenbush & Liu, 2000; Pituch et al., 2005). Hence, the treatment variable X is conceptually a variable at the unit-level. Within each cluster, more than one treatment condition can be realized at the same time. Treatment assignment probabilities for individual units can depend on properties of the unit and the cluster, i.e., on the unit variable U, the unit-covariate Z, the cluster variable C and the cluster-covariate V. This includes aspects of the within-cluster distribution of the unit-covariate Z, such as the between-component Z_b or conditional variances $Var(Z|C)$ and also the cluster-specific treatment propensities $P(X=j \mid C)$. Conceptually, this implies that all of these variables have to be considered as potential confounders.

In designs with treatment assignment at the cluster-level, clusters are assigned to treatment conditions as-a-whole (Donner & Klar, 2000; Murray, 1998). The treatment itself is administered to the complete cluster and, hence, the treatment variable is conceptually a variable at the cluster-level. Examples include job training programs, neighborhood interventions and different teaching techniques applied to classrooms. All units within a cluster receive the same treatment and only one treatment condition can be realized in a cluster at the same time. Treatment assignment probabilities can only depend on the cluster variable C and the cluster-covariate V and only these variables are potential confounders (see also Chapter 5). Again, elements of the within-cluster distributions of the unit-covariate Z such as the between-component Z_b or conditional variances $Var(Z|C)$ are among the confounders. However, the unit variable U, the unit-covariate Z or the within-cluster residual Z_w cannot influence the treatment assignment over and above the cluster-covariate V and, hence, cannot confound the effects of the treatment on the outcome variable Y.

3.2.2 Assignment of Units to Clusters

In line with the discussion of different assumptions about the assignment mechanism of units to clusters in the previous section, we distinguish three idealized assignment types: (1) *Pre-existing clusters* implying that the cluster variable C is a function of the unit variable U and no actual assignment of units to clusters takes place, (2) *randomized assignment* of units to clusters that implies stochastic independence of the unit variable U and the cluster variable C and (3) *non-randomized assignment* or self-selection of units to clusters with differing non-zero probabilities, including the case of stochastic independence of the unit variable U and the cluster variable C conditional on the unit-covariate Z (see also Roberts & Roberts, 2005, and the extensive discussion in the previous section).

3.2.3 Treatment Assignment Mechanism

The third dimension on which multilevel designs can be categorized has two levels: (1) *Experimental designs* and (2) *quasi-experimental designs* are distinguished by the knowledge and control of the treatment assignment mechanism (Shadish et al., 2002; Steyer et al., 2009). We will discuss further differentiations and the range of application of different design types in combination with the other dimensions of the taxonomy separately for both types of designs.

Experimental Designs

Experimental designs are designs in which the treatment assignment probabilities are known and under the control of the experimenter (Shadish et al., 2002). The classical example are *randomized experiments* in which units or clusters are assigned to treatment conditions with equal probabilities. *Conditionally randomized designs* are designs in which treatment assignment of units or clusters is randomized conditional on covariates.

Randomized Experiments. The classical case for experimental designs are *randomized experiments*, designs in which the individual treatment probabilities are equal for all units

$$P(X{=}j \mid U{=}u) = P(X{=}j) \quad \text{for all values } u \text{ of } U \text{ and } j \text{ of } X. \tag{3.5}$$

Randomization can be implemented for designs with treatment assignment at the unit-level as well as for designs with treatment assignment at the cluster-level. It can easily be implemented for designs with pre-existing clusters and designs with assignment of units to clusters. If units are randomly assigned to treatment conditions, all units have the same probability of being assigned to the treatment or control conditions, no matter which cluster the units belong to or have been assigned to. Randomization of units to treatment conditions also implies stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ as defined in Equation (2.59) (see Steyer et al., 2009, for the proof), that is a sufficient condition for unbiasedness of the treatment regression $E(Y \mid X)$.

For designs with treatment assignment at the cluster-level, each cluster c has the same probability of being assigned to the treatment condition

$$P(X{=}j \mid C{=}c) = P(X{=}j) \quad \text{for all values } c \text{ of } C \text{ and } j \text{ of } X. \tag{3.6}$$

Like randomized assignment of units to treatment conditions, randomized assignment of clusters to treatment conditions implies stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$. Since the unit variable U is measurable with respect to $\mathfrak{C}_X$, randomized assignment of clusters to treatment conditions implies stochastic independence of the unit variable U and the treatment variable X: The u-conditional treatment assignment probabilities $P(X{=}j \mid U{=}u)$ are equal to the unconditional treatment probability $P(X{=}j)$ — every unit u has the same probability of being assigned to each treatment condition. However, in samples, randomization at the cluster-level is likely to lead to less balance with respect to covariates between the treatment groups (see also Moerbeek et al., 2000; Raudenbush, 1997). Blocked or stratified randomization techniques are recommended alternatives to circumvent this problems and create equivalent treatment and control groups. (Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997; Raudenbush et al., 2007).

Conditionally Randomized Experiments. The second class of experimental designs are *conditionally randomized experiments*, also known as blocked or stratified randomized designs (Maxwell & Delaney, 2004; Raudenbush et al., 2007; Shadish et al., 2002): In these designs, treatment assignment is randomized given one or more covariates. In multilevel designs, different subtypes of conditionally randomized expe-

riments can be implemented depending on the covariates that are considered for randomization and the level of treatment assignment. Most of the designs — or more specifically all designs that include the unit-covariate Z among the variables conditional upon which randomization takes place — can only be meaningfully implemented for treatment assignment at the unit-level. We will now briefly review conditionally randomized between-group multilevel designs.

We begin with designs that use random assignment conditional on the unit-covariate Z. In this case, the treatment assignment probabilities are constant given a value of the unit-covariate Z, which directly means conditional independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ given the unit-covariate Z [as defined in Equation (2.60)]:

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid Z) \quad \text{for all values } j \text{ of } X. \tag{3.7}$$

Since conditional independence of X and $\mathfrak{C}_X$ implies unbiasedness of the unit-covariate-treatment regression $E(Y \mid X, Z)$, experimental multilevel designs with random assignment conditional on the unit-covariate Z also imply this condition. Obviously, random assignment to treatment conditions conditional on the values of the unit-covariate Z is only possible for designs with treatment assignment at the unit-level and not for designs with treatment assignment at the cluster-level. The same holds for randomized assignment conditional on values of the cluster variable C. In this case, all units within a cluster have the same probability of being assigned to either treatment condition, but the treatment assignment probabilities can differ between clusters. This implies conditional stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ given the cluster variable C [as defined in Equation (2.60)]:

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid C) \quad \text{for all values } j \text{ of } X. \tag{3.8}$$

Randomized assignment of whole clusters to treatment conditions with cluster-specific, but constant, treatment probabilities $P(X{=}j \mid C{=}c)$ would be possible, but would still mean that the primary units of assignment — the clusters — could differ with respect to their treatment assignment probabilities and would not remove the possible confounding due to cluster-covariates V.

Designs with randomized treatment assignment conditional on the cluster-covariate

V are the only class of conditionally randomized experiments that can be implemented for designs with treatment assignment the unit-level and as well as for designs with treatment assignment at the cluster-level. In these designs, the assignment probabilities are constant given values of the cluster-covariate V: In designs with treatment assignment at the unit-level, all units have the same probabilities of being assigned to either treatment or control conditions given the cluster-covariate V (including the between-component Z_b); in designs with treatment assignment at the cluster-level, all clusters have the same probabilities of being assigned to either treatment or control conditions given the cluster-covariate V. No matter the level of treatment assignment, conditional randomization on the cluster-covariate V guarantees conditional stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ given the cluster-covariate V

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid V) \quad \text{for all values } j \text{ of } X. \tag{3.9}$$

Again, this implies unbiasedness of $E(Y|X, V)$. We are implicitly including the between-component Z_b among the cluster-covariates V, since it is a function of the cluster variable C. However, in applications, conditional randomization given Z_b is hard to implement: The values of this variable are in general unobserved and only approximated by the empirical cluster means that are fallible measures of Z_b. Randomization conditional upon Z_b can only be implemented in designs with pre-existing clusters in which the unit-covariate can be assessed for all members of a cluster and would additionally require that the unit-covariate Z is measured without error. We will return to designs with randomization conditional on cluster-covariates V in Chapter 5.

Two additional classes of randomized designs are important with respect to the specification of adjustment models for designs with treatment assignment at the unit-level: These are designs in which assignment to treatment and control condition is conditionally random on combinations of values of the cluster variable C and of the unit-covariate Z, or less generally where assignment is conditionally random on combinations of values of the cluster-covariate V and the unit-covariate Z. In designs in which assignment to treatment conditions is randomized conditional upon the cluster variable C and the unit-covariate Z, each unit with a specific value of the unit-covariate z in cluster c has the same probabilities of being assigned to treatment or control conditions. This assignment scheme guarantees conditional stochastic independence of the treatment variable

X and the confounder σ-algebra $\mathfrak{C}_X$ given the unit-covariate Z and the cluster variable C:

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid Z, C) \quad \text{for all values } j \text{ of } X. \tag{3.10}$$

Similarly, in designs in which treatment assignment is conditionally randomized given the unit-covariate Z and the cluster-covariate V, i.e., in designs in which units with a value z of the unit-covariate and in a cluster with a value v of the cluster-covariate have equal probabilities of being assigned to treatment groups, the treatment variable X is conditionally independent of the confounder σ-algebra $\mathfrak{C}_X$ given the unit-covariate Z and the cluster-covariate V:

$$P(X{=}j \mid \mathfrak{C}_X) = P(X{=}j \mid Z, V) \quad \text{for all values } j \text{ of } X. \tag{3.11}$$

This independence includes stochastic independence of the cluster variable C and the treatment variable X. We will return to these two designs when we discuss adjustment models for designs with treatment assignment at the unit-level in Chapter 4.

By guaranteeing stochastic independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ unconditionally or conditionally, all experimental designs imply unbiasedness and unconfoundedness — once again either unconditionally in the case of randomized experiments and conditionally on the covariates which were used as stratification variables in conditionally randomized designs. Consequently, the mean differences between the treatment and control groups are unbiased estimators of the corresponding average causal effects in unconditionally randomized experiments under the usual distributional assumptions. In conditionally randomized experiments, the conditional mean differences are unbiased estimators of the conditional causal effects and can be used to obtain an unbiased estimators of the (unconditional) average causal effect as we will further discuss in Chapter 4 for designs with treatment assignment at the unit-level and in Chapter 5 for designs with treatment assignment at the cluster-level.

Quasi-Experimental Designs

In contrast to experimental designs in which treatment assignment is under the control of the experimenter and treatment probabilities are known and can be chosen in advance, quasi-experimental designs are designs in which treatment assignment probabilities are unknown and not under the control of the experimenter. Quasi-experimental

designs can be further distinguished in designs with *other-selection* of experimental units to treatment conditions and designs with *self-selection* of experimental units to treatment conditions (Heinsman & Shadish, 1996; Shadish & Heinsman, 1997; Shadish & Luellen, 2005).

Other-Selection. In quasi-experimental designs with *other-selection*, the assignment to treatment conditions is under the control of forces that are external to the entities that are assigned to treatment conditions (Heinsman & Shadish, 1996; Shadish et al., 2002). In quasi-experimental designs with treatment assignment at the unit-level, specific examples for other-selection include decisions about special education needs for individual students by teachers, decisions about the retention of individual students at a grade-level by teacher boards or decisions about therapeutic needs made by physicians for their patients. In designs with treatment assignment at the cluster-level, examples for other-selection include decisions of principals about the teaching methods to be implemented in different classes or decisions of school boards about which aspects of school reforms are to be implemented for different schools.

Self-Selection. In quasi-experimental designs with *self-selection*, the assignment to treatment conditions is based on decisions internal to the units of assignment (students, students' parents, teachers selecting a teaching method based on their knowledge of the class, Heinsman & Shadish, 1996; Shadish et al., 2002). In quasi-experimental designs with treatment assignment at the unit-level, examples for self-selection include the choice whether or not to actually follow a treatment regime by individual patients or the decision to attend voluntary afternoon classes or science clubs by students. In quasi-experimental designs with treatment assignment at the cluster-level, examples for self-selection include the selection of teaching methods for a class made jointly by students and teacher or the decisions to get external supervision made by working groups. In practice, the distinction between self- and other-selection in multilevel designs is not always as clear cut as it is for singlelevel designs (Shadish, 2000), especially when designs with treatment assignment at the cluster-level are considered. In these designs, decisions about treatment assignment will be almost always influenced by considerations of units within the clusters and external entities such as administrators or funding agencies.

Both types of quasi-experimental designs, no matter whether they rely on other- or

self-selection of experimental entities to treatment conditions, require assumptions of conditional unbiasedness and the careful selection of covariates at the unit- as well as at the cluster-level that influence both the treatment assignment and the outcome variable Y. In contrast to randomized designs, it is plausible to include the between-component Z_b among the cluster-covariates in quasi-experimental designs. It is more likely that self- or other-selection to treatment conditions depends on the true values of Z_b (e.g., the average socio-economic status in a school, or the average intelligence in a class) and not on the fallible observed cluster-means of Z assessed in an application. In contrast to conditionally randomized experimental designs, conditional unbiasedness is not guaranteed to hold in quasi-experimental designs. The correct analysis of quasi-experimental designs requires that all covariates on the unit- and cluster-level that guide the selection process to treatment conditions and influence the outcome variable Y are considered in order to estimate the average causal effect and approximate the results of experimental designs with quasi-experiments. The appropriate models for the estimation of average causal effects from multilevel quasi-experiments include the generalized ANCOVA introduced in Chapter 4 for designs with treatment assignment at the unit-level and in Chapter 5 for designs with treatment assignment at the cluster-level. Methods based on the estimated treatment propensities (e.g., Rosenbaum & Rubin, 1983, 1984, 1985) are an alternative, but have not been formally developed for multilevel designs. Meta-analytical studies of singlelevel designs (Heinsman & Shadish, 1996; Shadish, 2000) indicate that mechanisms governing the assignment process and the relevant covariates are more easily identified for designs with other-selection: Results from well-planned and analyzed quasi-experimental studies with other-selection closely mirror results from randomized experiments in the same subject domain.

3.3 Conclusion

In Chapter 2, we introduced the general theory of average causal effects, two multilevel single-unit trials to capture the structure of between-group multilevel designs and defined various causal effects, prima-facie effects, unbiasedness and some of its sufficient conditions on the causality space that formalizes the multilevel random experiment. In this chapter, we first discussed one big challenge to causal inference in multilevel designs — violations of SUTVA — and its relation to the assignment of units to clusters,

before presenting a taxonomy of multilevel designs in light of the preceding discussions. In this section that wraps up the chapter, we will summarize and discuss the prospects of causal inference for between-group multilevel designs.

The main insight from the general theory of causal effects (Steyer et al., 2009) is that causal inference can meaningfully refer only to the underlying probability and causality space, the distributions defined therein and the concepts defined thereupon. All changes in the underlying probability space, e.g., with respect to the population of units or clusters, to the probability of events or the distributions of random variables may also result in changes in the values of the derived quantities such as average causal effects. The scope of causal inferences is thus inseparably tied to the correct representation of the empirical phenomenon in the underlying multilevel random experiment: Without further assumptions, causal inferences can logically apply only to the population of clusters, the population of units and the assignment process of units to clusters the characterizes the random experiment. Inferences and generalizations above and beyond the random experiment, e.g., to other populations, time points or changes in underlying distributions, will always have to rely on additional assumptions. Changes in the causally relevant distributions and parameters in-between repetitions of the single-unit trial are especially critical: Estimation of average causal effects from finite samples requires a uniquely defined probability space as reference. If this probability space along with its distributions and regressions changes in the causal parameters, all inferences with respect to these parameters become meaningless.

As we have discussed, violations of SUTVA are widely discussed as challenges to valid causal inferences in multilevel designs. In contrast to previous accounts of causality in multilevel designs, the general theory of causal effects (Steyer et al., 2009) — by stratifying in its effect definition on the confounder σ-algebra $\mathfrak{C}_X$ — is well equipped to handle the complexities of multilevel designs. Thereby, true-effect variables τ_j always take all random variables that are measurable with respect to this σ-algebra into account. Additional interference effects between units within the same cluster or within and between treatment conditions can violate the stability and independence assumptions made in repeated single-unit trials and can challenge the validity of causal inferences in some designs. In designs with assignment of units to clusters, causal inference requires that the effects of interferences between units within the same cluster on the outcome variable Y are captured fully by contextual variables measurable with respect to $\mathfrak{C}_X$. Interference effects are unproblematic in multilevel designs with pre-existing

clusters and assignment at the cluster-level, since they are captured by the cluster variable C. However, average causal effects estimated from these designs can never be generalized to other allocations of units to clusters without the additional assumption that similar mechanisms will be present in the changed population as well. In all designs with treatment assignment at the unit-level, the additional and standard assumption of no interference or spill-over effects between the treated and the untreated units over and above the cluster-specific treatment propensities must hold. The development of analytical methods that can deal with additional SUTVA violations, e.g., by developing non-parametric bounds for treatment effects (Manski, 1995; Rosenbaum, 2007) or by using instrumental variable approaches (Sobel, 2006), is a fruitful area of further research for multilevel designs.

In the following two chapters, we will turn to the identification of the average causal effect in conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level (Chapter 4) and at the cluster-level (Chapter 5). In these two chapters, we will discuss another big challenge to causal inference: The correct specification of the regressions underlying the adjustment model. Additionally, we will further deliberate the choice of the correct statistical model to estimate the average causal effects and test various statistical implementations of the generalized ANCOVA for multilevel designs in two simulation studies. The third big challenge to causal inference in general and in multilevel designs — the choice of the appropriate covariates — will receive less attention in the course of this thesis. Basically, we will rely on conditional unbiasedness as the weakest causality criterion without further discussing indirect tests for unbiasedness. Although first approaches to implement tests of conditional unconfoundedness exist for singlelevel designs, they are not well understood yet and their application to multilevel designs is beyond the scope of this thesis.

4 Average Causal Effects for Treatment Assignment at the Unit-Level

This chapter discusses the analysis of causal effects in multilevel designs with treatment assignment at the unit-level. Since the analysis of randomized designs is well-understood (e.g., Moerbeek et al., 2000; Raudenbush & Liu, 2000) and average causal effects are identified with prima-facie effects in multisite randomized trials, we will focus on designs in which unbiasedness of the treatment regression $E(Y \mid X)$ does not hold. Specifically, we will develop an adjustment procedure that extends the generalized analysis of covariance (ANCOVA) introduced by Steyer et al. (2009) to conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level. The identification of average causal effects in these designs is not as well understood as causal inference for unconditionally randomized multisite designs. Conventional accounts of the multilevel ANCOVA for designs with treatment assignment at the unit-level (Moerbeek et al., 2001; Raudenbush & Liu, 2000; Seltzer, 2004; Pituch, 2001; Plewis & Hurry, 1998) do not identify the average causal effect in the presence of interactions between the treatment variable and the covariates.

The chapter is structured as follows: We first identify average causal effects for general conditional effect functions. Next, we turn to linear effect functions and develop the generalized ANCOVA for non-randomized multisite designs. We show how the average causal effect can be identified as a non-linear function of the parameters of a multiple linear regression. Then, we report the results of a simulation study that compared the finite sample performance of several statistical implementations under the null hypothesis of no average causal effect in a design with a unit-covariate Z and the corresponding between-component Z_b. Finally, we illustrate these implementations with an empiri-

cal example from the National Educational Longitudinal Study of 1988 (NELS:1988, Curtin, Ingels, Wu, Heuer, & Owings, 2002). Throughout this chapter, we will refer to the multilevel causality space $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ introduced in Chapter 2. Our discussion pertains to designs with assignment of units to clusters as well as to designs that use pre-existing clusters. In line with the discussions in Chapter 3, we assume that there are no additional spill-over effects between treated and untreated units within a cluster.

4.1 Adjustment Models

In the following section, we will develop the generalized ANCOVA (Steyer et al., 2009) for multilevel designs in which units are assigned to treatment conditions with differing probabilities, more specifically for conditionally randomized or quasi-experimental designs. In these designs, the mean differences of the outcome variable between treatment conditions, in general, do no longer identify the average causal effects.

In the remainder of the section, we consider two classes of conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level: First, we discuss designs, in which unbiasedness of the unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ holds [for the definition see Equation (2.57)]: In experimental designs, unbiasedness of $E(Y \mid X, Z, C)$ is implied, if all units with a value z of the unit-covariate Z within a cluster c have the same probabilities of being assigned to treatment and control conditions. In quasi-experimental designs, unbiasedness of $E(Y \mid X, Z, C)$ does not hold automatically, but requires the inclusion of all unit-covariates Z that influence the treatment probabilities and the outcome variable Y in addition to the cluster variable C. Next, we discuss designs in which the stronger condition of unbiasedness of the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V, Z_b)$ holds [for the definition see Equation (2.58)]. Again, in experimental designs, unbiasedness of $E(Y \mid X, Z, V, Z_b)$ is implied by assigning all units with a value z of the unit-covariate Z in clusters characterized by cluster-covariates $V{=}v$ and $Z_b{=}z_b$ with the same probabilities to treatment and control conditions. In quasi-experimental designs, unbiasedness of $E(Y \mid X, Z, V, Z_b)$ does not hold automatically, but requires the inclusion of all covariates Z, V and Z_b that influence both the treatment assignment probabilities and the outcome variable Y.

In the next section, we will show how the two unbiasedness assumptions allow the identification of average causal effects ACE_{jk} with empirically estimable quantities. We will first develop the adjustment models for general effect functions and then turn to effect functions that are linear in the covariates and their product variables.

4.1.1 General Effect Functions

In the following section, we will develop the adjustment model for general effect functions. We will start with designs in which unbiasedness of $E(Y|X,Z,C)$ holds and cover designs with unbiasedness of $E(Y \mid X,Z,V,Z_b)$ thereafter.

Unbiasedness of $E(Y \mid X,Z,C)$

We start our discussion of the identification of the average causal effect ACE_{jk} in non-randomized multilevel designs with treatment assignment at the unit-level with designs in which the unit-covariate-cluster regression $E(Y \mid X,Z,C)$ is unbiased. Unbiasedness of $E(Y \mid X,Z,C)$ has been defined in Equation (2.57) — for convenience, this definition is repeated here:

$$E^{\circ}_{X=j}(Y \mid Z,C) = E(\tau_j \mid Z,C) \quad \text{a.s. for all values } j \text{ of } X.$$

If the $(J+1)$-valued treatment variable X is represented with J dummy variables $I_{X=j}$ with values 0 and 1, one for each treatment condition using the control condition as reference, $E(Y \mid X,Z,C)$ can always be written as

$$E(Y \mid X,Z,C) = g_0(Z,C) + g_1(Z,C) \cdot I_{X=1} + \ldots + g_J(Z,C) \cdot I_{X=J}. \tag{4.1}$$

Without further assumptions, the function $g_0(Z,C)$ is the regression $E_{X=0}(Y \mid Z,C)$ of the outcome variable Y on the unit-covariate Z and the cluster variable C in the control group. The functions $g_j(Z,C)$ are the conditional prima-facie effect functions $PFE_{j0\,;\,Z,C}$ whose values are the conditional prima-facie effects of treatment condition j compared to the control condition for all combinations of the unit-covariate Z and the cluster variable C as defined in Equation (2.51):

$$g_j(Z,C) = E^{\circ}_{X=j}(Y \mid Z,C) - E^{\circ}_{X=0}(Y \mid Z,C) = PFE_{j0\,;\,Z,C}. \tag{4.2}$$

If the unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ is unbiased, the values of the $g_j(Z, C)$-functions are not only prima-facie effects, but also conditional causal effects. Hence, the $g_j(Z, C)$-functions are equal to the conditional causal effect functions $CCE_{j0;Z,C}$ as defined in Equation (2.36):

$$g_j(Z, C) = E(\tau_j \mid Z, C) - E(\tau_0 \mid Z, C) = CCE_{j0;Z,C}. \tag{4.3}$$

As shown in Equation (2.37), the expected value of the conditional causal effect function $E(CCE_{jk;Z,C})$ is equal to the average causal effect ACE_{jk}. Since the conditional causal effect function $CCE_{j0;Z,C}$ is — under unbiasedness of $E(Y \mid X, Z, C)$ — equal to $g_j(Z, C)$; its expected value $E[g_j(Z, C)]$ is equal to the average causal effect ACE_{j0}:

$$ACE_{j0} = E(CCE_{j0;Z,C}) = E[g_j(Z, C)]. \tag{4.4}$$

This equality can be used to identify average causal effects in non-randomized multilevel designs with treatment assignment at the level of the individual unit. In order to do this, the functional forms of $g_0(Z, C)$ and $g_j(Z, C)$ have to be specified and their parameters estimated or alternatively modeled non-parametrically (Steyer et al., 2009).

Unbiasedness of $E(Y \mid X, Z, V, Z_b)$

We are now turning to the identification of the average causal effect ACE_{jk} in designs with an unbiased unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V, Z_b)$. Unbiasedness of $E(Y \mid X, Z, V, Z_b)$ had been defined in Equation (2.58) — this definition is repeated here explicitly including the between-component Z_b among the random variables:

$$E^{\circ}_{X=j}(Y \mid Z, V, Z_b) = E(\tau_j \mid Z, V, Z_b) \quad \text{a.s. for all values of } X.$$

If J dummy variables $I_{X=j}$ with values 0 and 1 are used to represent the $(J + 1)$-valued treatment variable X, using the control condition as reference, the regression $E(Y \mid X, Z, V, Z_b)$ can always be written as

$$E(Y \mid X, Z, V, Z_b) = g_0(Z, V, Z_b) + g_1(Z, V, Z_b) \cdot I_{X=1} + \ldots + g_J(Z, V, Z_b) \cdot I_{X=J}. \tag{4.5}$$

Again, without further assumptions, the function $g_0(Z, V, Z_b)$ is always the regression

$E_{X=0}(Y \mid Z, V, Z_b)$ of the outcome variable Y on the unit-covariate Z, the cluster-covariate V and the between-component Z_b in the control group. The functions $g_j(Z, V, Z_b)$ are the conditional prima-facie effect functions $PFE_{j0\,;\,Z,\,V,\,Z_b}$ whose values are the conditional prima-facie effects of treatment condition j compared to the control condition for all combinations of the covariates Z, V and Z_b as defined in Equation (2.52):

$$g_j(Z, V, Z_b) = E^{\circ}_{X=j}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b) = PFE_{j0\,;\,Z,\,V,\,Z_b}. \tag{4.6}$$

Under unbiasedness of $E(Y \mid X, Z, V, Z_b)$, the values of the $g_j(Z, V, Z_b)$-functions are not only prima-facie effects, but also conditional causal effects and the $g_j(Z, V, Z_b)$-functions are equal to the conditional causal effect functions $CCE_{j0\,;\,Z,\,V,\,Z_b}$ as defined in Equation (2.40):

$$g_j(Z, V, Z_b) = E(\tau_j \mid Z, V, Z_b) - E(\tau_0 \mid Z, V, Z_b) = CCE_{j0\,;\,Z,\,V,\,Z_b}. \tag{4.7}$$

As shown in Equation (2.42), the expected value of the conditional causal effect function $E(CCE_{jk\,;\,Z,V,Z_b})$ is equal to the average causal effect ACE_{jk}. Since, under unbiasedness of $E(Y \mid X, Z, V, Z_b)$, the conditional causal effect function $CCE_{j0\,;\,Z,\,V,\,Z_b}$ is equal to $g_j(Z, V, Z_b)$, its expected value $E\big[g_j(Z, V, Z_b)\big]$ is equal to the average causal effect ACE_{j0}:

$$ACE_{j0} = E(CCE_{j0\,;\,Z,\,V,\,Z_b}) = E\big[g_j(Z, V, Z_b)\big]. \tag{4.8}$$

This equality can be used to identify average causal effects in non-randomized multilevel designs with treatment assignment at the level of the individual unit. Again, in applications, the functional forms of $g_0(Z, V, Z_b)$ and $g_j(Z, V, Z_b)$ have to be parametrized and estimated or modeled non-parametrically.

4.1.2 Linear Effect Functions

The unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ and the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V, Z_b)$ were introduced in Equations (4.1) and (4.5) without further assumptions about the form of the respective intercept functions g_0 and effect functions g_j. If the adjustment methods are used in applications, the functional forms of g_0 and g_j must be explicitly chosen or estimated with other, e.g., non-parametric statistical methods. If an explicit specification of the g_0 and g_j is chosen,

the validity of causal inferences depends on the correct specification of the regression of the outcome variable Y on the set of considered covariates in each treatment group. Any misspecification of these regressions can result in a severe bias of the estimated average causal effect (see also, D. J. Bauer & Cai, 2008; Kang & Schafer, 2006).

In the remainder of this section, we will develop the generalized ANCOVA (Kröhne, 2009; Steyer et al., 2009) for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level using linear intercept and effect functions (see also, Kröhne, 2009, for a similar formulation with regard to the parametrization of the regressions of the outcome variable Y on the covariate Z for each treatment group for the generalized ANCOVA in singlelevel designs). We will start with designs that lead to an unbiased unit-covariate-cluster-treatment regression $E(Y|X,Z,C)$ and discuss designs with an unbiased unit-covariate-cluster-covariate-treatment regression $E(Y \mid X,Z,V,Z_b)$ thereafter. If a linear parametrization for the intercept and effect functions is chosen, the validity of causal inferences is contingent not only on the assumption of unbiasedness of the respective regressions, but also on the tenability of the linearity assumptions for the functions g_0 and g_j.

The generalized ANCOVA, introduced in the next section, extends and consolidates the different versions of multilevel ANCOVA models that have been proposed for nonrandomized multilevel designs with treatment assignment at the unit-level (Moerbeek et al., 2001; Raudenbush & Liu, 2000; Seltzer, 2004; Pituch, 2001; Plewis & Hurry, 1998) in the following ways:

1. In line with Moerbeek et al. (2001), it explicitly acknowledges the decomposition of the unit-covariate Z into the within-component Z_w and the between-component Z_b.

2. Similar to Pituch (2001), Plewis and Hurry (1998), Raudenbush and Liu (2000) and Seltzer (2004), it includes interactions between the covariates Z, V, Z_b and the treatment variable X, but extends the aforementioned approaches by explicitly identifying the average treatment effect in the presence of interactions and not only conditional treatment effects (see also Flory, 2008; Kröhne, 2009; Nagengast, 2006; Rogosa, 1980).

3. Finally, it is embedded in an explicit theory of causality and uses covariates to identify the average causal effect ACE_{jk} in conditionally-randomized and quasi-

experimental designs — in contrast to Moerbeek et al. (2001) and Raudenbush and Liu (2000) who discussed covariates only as a means to improve precision in randomized designs.

Unbiasedness of $E(Y \mid X, Z, C)$

Our discussion of the adjustment model for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level that lead to unbiasedness of $E(Y \mid X, Z, C)$ will be confined to a binary treatment variable X, representing a treatment and a control condition. To further simplify the derivations, we will only consider one unit-covariate Z and the cluster variable C as covariates. Generalizations to more than two treatment conditions and more unit-covariates are straightforward, but require further assumptions about the interactions between the covariates.

If the regression $E(Y \mid X, Z, C)$ is conditionally linear in the treatment variable X, the unit-covariate Z and their product variable given the cluster variable C, it can be written as:

$$E(Y \mid X, Z, C) = f_0(C) + f_1(C) \cdot Z + \big[f_2(C) + f_3(C) \cdot Z\big] \cdot X. \tag{4.9}$$

Equation (4.9) shows that the $(C{=}c)$-conditional regressions $E_{C=c}(Y \mid X, Z)$ are linear in the treatment variable X, the unit-covariate Z and their product variable with cluster-specific regression parameters $f_0(C{=}c)$ to $f_3(C{=}c)$. The $(C{=}c)$-conditional regressions are also referred to as within-cluster regressions or level-1-equations in conventional multilevel modeling terminology (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Equation (4.9) is not restrictive, if the unit-covariate Z is a dichotomous variable. In this case the functions $f_i(C)$ can be modeled using indicator variables $I_{C=c}$ for the clusters $C{=}c$, the unit-covariate Z and their products. If Z is a discrete random variable with a small number of possible values, indicator variables $I_{Z=z}$ can be used to obtain a saturated parametrization of Equation (4.9). If Z is a continuous variable (or a many-valued discrete random variable, that cannot be reasonably represented with indicator variables), however, the linearity of the $(C{=}c)$-conditional regressions $E_{C=c}(Y \mid X, Z)$ postulated in Equation (4.9) can be wrong.

The outcome variable Y can be decomposed into the regression $E(Y \mid X, Z, C)$ and its residual $\varepsilon \equiv Y - E(Y \mid X, Z, C)$:

$$Y = f_0(C) + f_1(C) \cdot Z + \big[f_2(C) + f_3(C) \cdot Z\big] \cdot X + \varepsilon. \tag{4.10}$$

The residual ε has all properties of a residual of a multiple linear regression, most notably its expected value and its regression on the regressors is zero. The last property only holds, if the actual regression of the outcome variable Y on X, Z and C is c-conditionally linear in X, Z and their product. Otherwise, only the properties of the residual of a linear ordinary least-squares regression $Q(Y \mid X, Z, C)$ hold (e.g., Kutner, Nachtsheim, Neter, & Li, 2005; Rechner & Schaalje, 2007, Chapter 8). No further assumptions about the distribution of the residual ε are made at this point.

Next, we derive the conditional effect function $g_1(Z, C)$ from $E(Y \mid X, Z, C)$ in Equation (4.9). In Equation (4.2), we noted that $g_1(Z, C)$ is equal to the difference between the extensions of the conditional regressions $E_{X=1}(Y \mid Z, C)$ and $E_{X=0}(Y \mid Z, C)$. Rewriting Equation (4.9), these two regressions have the following form:

$$E_{X=0}(Y \mid Z, C) = f_0(C) + f_1(C) \cdot Z, \tag{4.11}$$

$$E_{X=1}(Y \mid Z, C) = [f_0(C) + f_2(C)] + [f_1(C) + f_3(C)] \cdot Z. \tag{4.12}$$

Equation (4.11) describes the regression $E_{X=0}(Y \mid Z, C)$ of the outcome variable Y on the unit-covariate Z and the cluster variable C in the control condition and is thus equal to the intercept function $g_0(Z, C)$. The conditional effect function $g_1(Z, C)$ is obtained by subtracting the extension of Equation (4.11) from the extension of Equation (4.12):

$$g_1(Z, C) = E^{\circ}_{X=1}(Y \mid Z, C) - E^{\circ}_{X=0}(Y \mid Z, C) \tag{4.13}$$

$$= f_2(C) + f_3(C) \cdot Z. \tag{4.14}$$

To obtain the average causal effect ACE_{10}, the expected value of the conditional effect function $g_1(Z, C)$ has to be taken [see Equation (4.4)]:

$$E[g_1(Z, C)] = E[f_2(C) + f_3(C) \cdot Z] \tag{4.15}$$

$$= E[f_2(C)] + E[f_3(C) \cdot Z]. \tag{4.16}$$

At this point, further assumptions about the functions $f_2(C)$ and $f_3(C)$ are necessary to identify the average causal effect ACE_{10}. While the expected values $E[f_2(C)]$ can be easily identified and estimated, e.g., with a hierarchical linear model (Raudenbush & Bryk, 2002), the expected value of the product term $E[f_3(C) \cdot Z]$ is harder to obtain, since it depends both on the covariance and the expected values of $f_3(C)$ and the unit-

covariate Z:

$$E[f_3(C) \cdot Z] = Cov[f_3(C), Z] + E[f_3(C)] \cdot E(Z). \tag{4.17}$$

If the covariance $Cov[f_3(C), Z]$ is equal to zero, the expected value $E[f_3(C) \cdot Z]$ is identified by the product of the expected values $E[f_3(C)]$ and $E(Z)$. This will be the case in designs in which units are randomly assigned to clusters. In this case $E(Z \mid C) = E(Z)$ which implies that $Cov[f_3(C), Z] = 0$. In designs with pre-existing clusters and non-randomized assignment or self-selection to clusters, the covariance $Cov[f_3(C), Z]$ will usually not be equal to zero and the cluster-specific effects of the product of the treatment variable X and the unit-covariate Z can covary with the unit-covariate Z. In these designs, the average causal effect ACE_{10} is identified, but practical estimation can be difficult, since statistical multilevel models require no correlation between error terms and predictor variables within and between levels (see also, Kim & Frees, 2006, 2007). We will not consider designs with unbiasedness of $E(Y \mid X, Z, C)$ further. Instead, we will focus on designs with unbiasedness of $E(Y \mid X, Z, V, Z_b)$ and develop and study the generalized ANCOVA for these designs in more detail.

Unbiasedness of $E(Y \mid X, Z, V, Z_b)$

As in the previous section, our discussion of the adjustment model for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level and unbiasedness of $E(Y \mid X, Z, V, Z_b)$ will be confined to a binary treatment variable X, representing a treatment and a control condition. Again, we will only consider one unit-covariate Z, one cluster-covariate V, the between-component Z_b and their products. Generalizations to more than two treatment conditions and more covariates are straightforward, but require further assumptions about the interactions between the covariates to be considered.

There are two equivalent ways to parametrize the regression $E(Y \mid X, Z, V, Z_b)$: (1) Either with the raw scores of the unit-covariate Z or (2) with the within-component Z_w [as defined in Equation (2.3)]. Both parametrizations include identical information about the regression of the outcome variable on the treatment variable and the covariates since the within-component Z_w is defined as the difference between the raw scores of the unit-covariate Z and the between-component Z_b (see also, Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995), but the interpretation of regression parameters differ. For the following derivations, the parametrization using the within-component Z_w is used, as it

results in an easier identifier of the average causal effect. At the end of the section, we will briefly discuss the alternative parametrization of $E(Y \mid X, Z, V, Z_b)$ that uses the raw scores of the unit-covariate Z.

If the regression $E(Y \mid X, Z, V, Z_b)$ is linear in the treatment variable X, the within-component Z_w, the between-component Z_b, the cluster-covariate V and their products, it can be written as:

$$\begin{aligned} E(Y \mid X, Z, V, Z_b) = {} & \gamma_{00} + \gamma_{01} Z_b + \gamma_{02} V + \gamma_{03} Z_b \cdot V \\ & + \left[\gamma_{04} + \gamma_{05} Z_b + \gamma_{06} V + \gamma_{07} Z_b \cdot V\right] \cdot Z_w \\ & + \left[\gamma_{10} + \gamma_{11} Z_b + \gamma_{12} V + \gamma_{13} Z_b \cdot V \right. \\ & \left. + \left[\gamma_{14} + \gamma_{15} Z_b + \gamma_{16} V + \gamma_{17} Z_b \cdot V\right] \cdot Z_w\right] \cdot X. \end{aligned} \tag{4.18}$$

The outcome variable Y can be decomposed into the regression $E(Y \mid X, Z, V, Z_b)$ and its residual $\varepsilon \equiv Y - E(Y \mid X, Z, V, Z_b)$:

$$Y = E(Y \mid X, Z, V, Z_b) + \varepsilon. \tag{4.19}$$

Again, all properties of a residual of a multiple regression apply to the residual ε, most notably its expected value and its regression on the regressors is zero. The last property only holds, if the actual regression of the outcome variable Y on the regressors X, Z_w, Z_b, V and their products is linear. Otherwise, only the properties of the residual of a linear ordinary least-squares regression $Q(Y \mid X, Z, V, Z_b)$ hold (e.g., Kutner et al., 2005; Rechner & Schaalje, 2007, Chapter 8). No further assumptions about the distribution of ε are made at this point. However, assumptions about the distribution of the residual and its multilevel structure distinguish the statistical models used to implement the generalized ANCOVA in the simulation study in Section 4.2.

Next, we derive the parameters of the conditional effect function $g_1(Z, V, Z_b)$ from the parameters of Equations (4.18). In Equation (4.6), we noted that $g_1(Z, V, Z_b)$ is equal to the difference between the extensions of the conditional regressions $E_{X=1}(Y \mid Z, V, Z_b)$ and $E_{X=0}(Y \mid Z, V, Z_b)$. Taking the parameters of the regression in Equation (4.18), these

two regressions have the following form:

$$\begin{aligned} E_{X=0}(Y \mid Z, V, Z_b) &= \gamma_{00} + \gamma_{01}Z_b + \gamma_{02}V + \gamma_{03}Z_b \cdot V \\ &\quad + [\gamma_{04} + \gamma_{05}Z_b + \gamma_{06}V + \gamma_{07}Z_b \cdot V] \cdot Z_w, \end{aligned} \tag{4.20}$$

$$\begin{aligned} E_{X=1}(Y \mid Z, V, Z_b) &= \gamma_{00} + \gamma_{10} + (\gamma_{01} + \gamma_{11})Z_b \\ &\quad + (\gamma_{02} + \gamma_{12})V + (\gamma_{03} + \gamma_{13})Z_b \cdot V \\ &\quad + [\gamma_{04} + \gamma_{14} + (\gamma_{05} + \gamma_{15})Z_b \\ &\quad + (\gamma_{06} + \gamma_{16})V + (\gamma_{07} + \gamma_{17})Z_b \cdot V] \cdot Z_w. \end{aligned} \tag{4.21}$$

Equation (4.20) describes the regression $E_{X=0}(Y \mid Z, V, Z_b)$ of the outcome variable Y on the covariates in the control condition and is thus equal to the intercept function $g_0(Z, V, Z_b)$. The conditional effect function $g_1(Z, V, Z_b)$ is obtained by subtracting the extension of Equation (4.20) from the extension of Equation (4.21):

$$\begin{aligned} g_1(Z, V, Z_b) &= E^{\circ}_{X=1}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b) \\ &= \gamma_{10} + \gamma_{11}Z_b + \gamma_{12}V + \gamma_{13}Z_b \cdot V \\ &\quad + [\gamma_{14} + \gamma_{15}Z_b + \gamma_{16}V + \gamma_{17}Z_b \cdot V] \cdot Z_w. \end{aligned} \tag{4.22}$$

To identify the average causal effect ACE_{10}, the expected value of Equation (4.22) has to be taken [see Equation (4.8)] and the algebraic rules for expected values (cf. H. Bauer, 1981) have to be applied to the resulting function:

$$\begin{aligned} E[g_1(Z, V, Z_b)] &= \gamma_{10} + \gamma_{11}E(Z_b) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V) \\ &\quad + \gamma_{14}E(Z_w) + \gamma_{15}E(Z_b \cdot Z_w) \\ &\quad + \gamma_{16}E(V \cdot Z_w) + \gamma_{17}E(Z_b \cdot V \cdot Z_w) \end{aligned} \tag{4.23}$$

$$= \gamma_{10} + \gamma_{11}E(Z) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V). \tag{4.24}$$

Equation (4.23) simplifies considerably to Equation (4.24). These simplifications are possible by taking into account that Z_w is the residual of the regression $E(Z \mid C)$ (see Section 2.3 for the discussion of this regression and its residual). By virtue of this relation, its expected value is equal to zero [$E(Z_w) = 0$] and it is regressively independent from all functions of its regressor C, making the expected values of its products with Z_b, V and $Z_b \cdot V$ also equal to zero [$E(Z_b \cdot Z_w) = 0$, $E(V \cdot Z_w) = 0$ and $E(Z_b \cdot V \cdot Z_w) = 0$],

at least if the cluster-covariate V is a function of the cluster variable C. In case of a cluster-covariate V that is not a function of C, the corresponding parameter γ_{16} of the interaction between Z_w and V remains in Equation (4.24).

Thus, the average causal effect ACE_{10} in experimental multilevel designs with conditional randomization or quasi-experimental multilevel designs with treatment assignment at the unit-level assuming linear intercept- and effect-functions, taking separate within- and between-effects of the unit-covariate Z into account and assuming unbiasedness of the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V, Z_b)$ is given by the following non-linear function of model parameters and expected values of the covariates and the corresponding product

$$ACE_{10} = \gamma_{10} + \gamma_{11}E(Z) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V). \tag{4.25}$$

As in the singlelevel generalized ANCOVA (Kröhne, 2009; Steyer et al., 2009), the ACE_{10} is identified as a non-linear function of regression parameters and expected values of the covariates and their products. If there are no interactions between the treatment variable X and the covariates, i.e., if the corresponding regression weights γ_{11}, γ_{12} and γ_{13} of the product terms of the treatment variable and the corresponding covariates are *all* equal to zero and, hence, the conditional effect function $g_1(Z, V, Z_b)$ is a constant, the average causal effect ACE_{10} is identified by the regression parameter γ_{10} of the treatment indicator variable. This is in line with presentations of the conventional multilevel ANCOVA in general (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and for designs with treatment assignment at the unit-level in particular (Moerbeek et al., 2000) where no interactions between the treatment variable and the covariates are included. It is well-known from singlelevel ANCOVA that — in the presence of non-zero interactions between the treatment and the covariates — the estimate from a model without interactions does not identify the average causal effect ACE_{10} in the presence of non-zero interaction effects, but only the conditional treatment effect at the point of highest precision [see Rogosa, 1980, Equations (18) and (19)]. Hence, in contrast to the generalized ANCOVA, the conventional multilevel ANCOVA does not identify the average causal effect ACE_{10} generally.

If the expected values of the covariates and their product — $E(Z)$, $E(V)$ and $E(Z_b \cdot V)$ — are equal to zero, the average causal effect ACE_{10} is also simply identified by the regression parameter γ_{10} of the treatment indicator variable. Centering covariates by

subtracting their empirical means from the observed values in applications before calculating the products with the treatment variable, makes their empirical mean equal to zero (Aiken & West, 1996; Kreft et al., 1995) — but not necessarily the observed mean of the product variable that also depends on the covariance $Cov(Z_b, V)$ that is not influenced by centering. While centering covariates around the empirical means identifies the average causal effect with a single regression weight — provided there are no three-way or higher-order interactions of the covariates and the treatment variable — analytical derivations (Chen, 2006) and simulation studies (Kröhne, 2009; Nagengast, 2006) for statistical implementations of singlelevel adjustment models show that standard errors of the *ACE* are underestimated, if covariates are stochastic predictors and not fixed by design. Centering can therefore not be recommended without qualifications for obtaining standard errors of the average causal effect for designs with stochastic covariates and observational studies.

In Equation (4.22), the conditional effect function $g_1(Z, V, Z_b)$ specifically provided for separate effects of both the between-component Z_b, the within-component Z_w and of product variables including these two variables. If the effects of Z_b and Z_w and their product variables with the cluster-covariate V do not differ, i.e., if $\gamma_{11}=\gamma_{14}$ and $\gamma_{13}=\gamma_{16}$, and there are no cross-level interactions, i.e., $\gamma_{15}=\gamma_{17}=0$, it is sufficient to model the raw scores of the unit-covariate Z directly. In this case, the corresponding regression weight of Z and of possible product variables of Z and the cluster-covariate V remain in the non-linear constraint. If either of these conditions is not met, the regression using that includes only the raw scores of the unit-covariate Z is misspecified and the resulting identifier of the ACE_{01} can be biased.

As mentioned above, the regression $E(Y \mid X, Z, V, Z_b)$ was parametrized using the within-component Z_w to derive and identify the ACE_{10} in the generalized ANCOVA while an alternative parametrization uses the raw scores of the unit-covariate Z. The two parametrizations of $E(Y|X, Z, V, Z_b)$ convey identical information with respect to the conditional expected values of the outcome variable Y (Enders & Tofighi, 2007; Kreft et al., 1995). However, the interpretation of model parameters and the resulting non-linear constraint differ. The specification using the within-cluster residual Z_w is to be preferred for models with interactions because the coefficients are easier interpretable as effects on different hierarchical levels and model estimation is more stable because the within-component Z_w is regressively independent of the between-component Z_b and, hence, their covariance is zero (Enders & Tofighi, 2007).

4.2 Simulation Study

In this section, we describe a large simulation study that compared the finite sample performance of several statistical implementations of the generalized ANCOVA with linear effect and intercept functions for non-randomized multilevel designs with treatment assignment at the unit-level. In the simulation, we compared several statistical models under an average causal effect of zero in the population to establish and test their adequacy under the null hypothesis. In order to simplify the simulation, we only considered a design with a unit-covariate Z and did not include a cluster-covariate V. However, the unit-covariate Z influenced both the treatment assignment and the outcome variable Y independently through its between-component Z_b and its within-component Z_w.

The section is structured as follows: We first introduce the data generation and assumptions made in repeating the single-unit trial. Next, we introduce the statistical models and research questions to be addressed in the simulation study. We then describe the design of the simulation study and report the results: The *ACE*-estimators were studied with respect to their bias and their relative efficiency, the standard errors of the *ACE*-estimators were studied with respect to their bias and the empirical type-1-error rate in tests of the null hypothesis.

4.2.1 Data Generation

The generalized ANCOVA was developed in the previous section with reference to the single-unit trial and the causality space introduced in Chapter 2 (see also, Steyer et al., 2009). Inferences from finite samples that are the focus of the simulation study require independent repetitions of a single-unit trial that are stable with respect to causal parameters and distributions and a statistical model to estimate the average causal effect and its standard error (see also the discussions in Chapter 3). In this section, we describe the central components of the repeated single-unit trial that was considered in the simulation study and discuss the resulting properties of the sampling model. A detailed description of the implementation of the data generation in R (R Development Core Team, 2008) is given in Appendix C.1.

In line with the theoretical concepts and data generation procedures in other simulation studies of the analysis of average causal effects with the generalized ANCOVA for singlelevel models (Kröhne, 2009), data was generated by considering the regressions

of the true-outcome variable τ_0 and the true-effect variable δ_{10} on the covariates Z and Z_b. Hence, τ_0 and δ_{10} were decomposed in the regressions $E(\tau_0 \,|\, Z, Z_b)$ and $E(\delta_{10} \,|\, Z, Z_b)$ and their respective residuals ε_0 and ε_{10}:

$$\tau_0 = E(\tau_0 \mid Z, Z_b) + \varepsilon_0, \tag{4.26}$$

$$\delta_{10} = E(\delta_{10} \mid Z, Z_b) + \varepsilon_{10}, \tag{4.27}$$

where $E(\delta_{10} \mid Z, Z_b)$ is the conditional effect function $CCE_{10;Z,Z_b}$. The presence of the residuals ε_0 and ε_1 violated conditional homogeneity as introduced in Section 2.6.2: The true-outcome variable τ_0 in the control condition and the true-effect variable δ_{10} were not constant given the covariates Z and Z_b. The residuals ε_0 and ε_1 were assumed to be further decomposable into cluster-specific components $r_{j;C}$ and unit-specific components $\upsilon_{j;U}$ to represent the effects of the cluster variable C and the unit variable U

$$\varepsilon_0 = r_{0;C} + \upsilon_{0;U} \tag{4.28}$$

$$\varepsilon_{10} = r_{10;C} + \upsilon_{10;U}. \tag{4.29}$$

The unit-specific components $\upsilon_{0;U}$ and $\upsilon_{10;U}$ are defined as the residuals of the regressions $E(\tau_0 \mid Z, Z_b, C)$ and $E(\delta_{10} \mid Z, Z_b, C)$ respectively. As residuals, their expected value is equal to zero and they are regressively independent of the regressors Z, Z_b and C. Furthermore, in the repeated single-unit trials of the simulation, their covariance $Cov\,(\upsilon_{0;U}, \upsilon_{10;U})$ was fixed to zero. The cluster-specific components $r_{0;C}$ and $r_{10;C}$ account for the multilevel structure of the repeated single-unit trials: They capture the residual influence of the cluster variable C on the true-outcome variable τ_0 in the control condition and the true-effect variable δ_{10} and are defined as residuals of the regressions $E\,[E(\tau_0 \mid Z, Z_b, C) \mid Z, Z_b]$ or $E\,[E(\delta_{10} \mid Z, Z_b, C) \mid Z, Z_b]$ respectively. Hence, their expected value is equal to zero and their regression on the regressors Z and Z_b is zero. Also, their covariance with the unit-specific residuals $\upsilon_{0;U}$ and $\upsilon_{10;U}$ is zero by definition. Additionally, their covariance $Cov\,(r_{0;C}, r_{10;C})$ was zero in the data generation procedure. If the variance of the residual ε_{10} is larger than zero, there will be residual variance heterogeneity between the treatment groups. Depending on whether this heterogeneity is due to $Var(r_{10;C}) > 0$ or $Var(\upsilon_{10;U}) > 0$, the heterogeneity is located at the unit- or at the cluster-level or due to both.

The decomposition of the residuals ε_0 and ε_{10} yielded the following residual intra-

class correlation coefficients *rICC*s in line with the general definition in Equation (2.11):

$$rICC(\tau_0 \mid Z, Z_b) = \frac{Var\left[E(\tau_0 \mid Z, Z_b, C) - E(\tau_0 \mid Z, Z_b)\right]}{Var\left[\tau_0 - E(\tau_0 \mid Z, Z_b)\right]} = \frac{Var(r_{0;C})}{Var(r_{0;C}) + Var(\upsilon_{0;U})}, \tag{4.30}$$

$$rICC(\delta_{10} \mid Z, Z_b) = \frac{Var\left[E(\delta_{10} \mid Z, Z_b, C) - E(\delta_{10} \mid Z, Z_b)\right]}{Var\left[\delta_{10} - E(\delta_{10} \mid Z, Z_b)\right]} = \frac{Var(r_{10;C})}{Var(r_{10;C}) + Var(\upsilon_{10;U})}. \tag{4.31}$$

If Equations (4.30) and (4.31) are both equal to zero, i.e., if the variances of the cluster-specific components $r_{0;C}$ and $r_{10;C}$ are both equal to zero, the estimated distribution of the parameter estimates from a conventional singlelevel regression model will be correct. If these variances are different from zero, the standard errors and covariance of parameter estimates will be underestimated, if the corresponding variance components of the outcome variable Y are not included in the statistical model (Moerbeek et al., 2000, 2001; Raudenbush & Liu, 2000; Snijders & Bosker, 1999).

The regressions $E(\tau_0 \mid Z, Z_b)$ and $E(\delta_{10} \mid Z, Z_b)$ were parameterized as linear functions of the within-component Z_w, the between-component Z_b and their product using the same labels for the regression coefficients as in Equation (4.18):

$$E(\tau_0 \mid Z, Z_b) = \gamma_{00} + \gamma_{01} \cdot Z_b + \gamma_{04} \cdot Z_w + \gamma_{05} \cdot Z_b \cdot Z_w, \tag{4.32}$$

$$E(\delta_{10} \mid Z, Z_b) = \gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w. \tag{4.33}$$

Both, Equations (4.32) and (4.33), used the true values z_b of the between-component Z_b, the (C=c)-conditional expected values $[E(Z \mid C{=}c)]$, and the residual of this regression, the within-component Z_w. However, the true values of Z_b were not available in the sample and had to be approximated by the cluster-means of Z and the empirical deviations from the cluster-means. If the regression weights of the between- and the within-component in Equations (4.32) and (4.33) are equal, i.e., if $\gamma_{01} = \gamma_{04}$ and $\gamma_{11} = \gamma_{14}$, and if there are no interactions between Z_b and Z_w, i.e., $\gamma_{05} = \gamma_{15} = 0$, or if the $ICC(Z)$=0, the multilevel decomposition of the unit-covariate does not have to be taken into account explicitly [see Equation (2.13)]. If at least one of these conditions is not fulfilled, a statistical model that only includes the unit-covariate Z as predictor will be misspecified and lead to a biased estimator of the *ACE*. The cluster variable C did not modify the effects of the within-component Z_w or the product variable $Z_b \cdot Z_w$, i.e.,

the parameters γ_{04}, γ_{05}, γ_{14} and γ_{15} were constant across clusters. The average causal effect is the expected value of the true-effect variable δ_{10} given in Equation (4.27) using the decomposition of the residual ε_{10} in Equation (4.29) and the parametrization in Equation (4.33):

$$ACE_{10} = E(\gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w + r_{10;C} + \upsilon_{10;U}), \tag{4.34}$$

$$= \gamma_{10} + \gamma_{11} \cdot E(Z). \tag{4.35}$$

Treatment assignment probabilities were determined by a logistic function that described the probability of units being assigned to each treatment condition as a function of the unit-covariate Z and the between-component Z_b:

$$P(X{=}1 \mid Z, Z_b) = \frac{exp\,(g_0 + g_1 \cdot Z_w + g_2 \cdot [Z_b - E(Z)])}{1 + exp\,(g_0 + g_1 \cdot Z_w + g_2 \cdot [Z_b - E(Z)])}. \tag{4.36}$$

Again the true values of the within-component Z_w and the between-component Z_b were used to parametrize Equation (4.36). This function guaranteed independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ given the unit-covariate Z and the cluster-component Z_b (i.e., $X \perp \mathfrak{C}_X \mid Z, Z_b$) as a sufficient condition for unbiasedness and unconfoundedness of $E(Y \mid X, Z, Z_b)$. If both parameters, g_1 and g_2, were equal to zero, treatment assignment probabilities would not depend on the between-component Z_b and the within-component Z_w and the data generation would represent a multisite randomized experiment with equal treatment probabilities in each cluster.

Each data set for the simulation study was generated by repeating a single-unit trial as described above with further assumptions about the distributions and parameters involved. In line with other simulation studies of the analysis of average causal effects and in order to represent quasi-experimental designs (Flory, 2008; Kröhne, 2009; Nagengast, 2006), the realized values of the unit-covariate Z in each sample were not fixed, but obtained by sampling from the unconditional distribution of the unit-covariate Z in each repetition of the single-unit trial. Hence, the realized values of Z varied from sample to sample and the unit-covariate and by implication the between-component Z_b and the within-component Z_w were stochastic predictors (Chen, 2006; Gatsonis & Sampson, 1989; Nagengast, 2006; Sampson, 1974; Shieh, 2006). In a similar vein, the use of the probabilistic assignment function, given in Equation (4.36), yielded samples that varied in the treatment group sizes depending on the realized values of the covariate and

the actual assignment of units to treatment conditions in the repeated single-unit trials. Hence, the treatment variable X also was a stochastic predictor (Nagengast, 2006). A detailed description of the parameters that were varied or kept constant within the simulation design is given in Section 4.2.3. The implementation of the data generation in R is explained and the corresponding parameter values of the data generation procedure are given in Appendix C.1.

Summarizing, the data generation routine resulted in four special properties of the sampling model that were represented to a different degree by the statistical models in the simulation study:

1. Due to repeated sampling from the unconditional distribution of the unit-covariate Z, its realized values varied from sample to sample. Thus, the unit-covariate Z and by implication the between-component Z_b and the within-component Z_w were stochastic predictors (Chen, 2006; Sampson, 1974; Shieh, 2006). The same logic applied to the realizations of the treatment variable X.

2. The conditional effect function $CCE_{10\,;\,Z,\,Z_b}$ varied independently with both the between-component Z_b and the within-component Z_w, as did the treatment assignment function. Thus, Z_b and Z_w are both relevant covariates to be considered separately in adjustment models, if $ICC(Z) > 0$ [see Equation (2.13)].

3. The regressions $E(\tau_0 \mid Z, Z_b)$ and $E(\delta_{10} \mid Z, Z_b)$ were specified using the actual values of the between-component Z_b. These values are only approximated by the empirical cluster means of Z that are fallible measures of $E(Z \mid C{=}c)$ in samples (Asparouhov & Muthén, 2006; Lüdtke et al., 2008). The average reliability of the cluster means is a function of the average cluster sizes and determines whether the estimated regression coefficients associated with the empirical cluster-means of Z will be biased [see Equation (2.14)].

4. Both $E(\tau_0 \mid Z, Z_b)$ and $E(\delta_{10} \mid Z, Z_b)$ were allowed to have residual intraclass correlation coefficients $rICC$ larger than zero, reflecting systematic influences of the cluster variable C on τ_0 and δ_{10} after taking the effects of the unit-covariate Z and the between-component Z_b into account. In addition, the variances of all residuals were not restricted to be equal, potentially resulting in variance heterogeneity of the outcome variable Y between treatment groups.

In the following section, we will discuss several statistical models and their abilities to deal with the properties of the data generation procedure.

4.2.2 Research Questions and Statistical Methods

The simulation study investigated the finite-sample performance of several statistical implementations of the generalized ANCOVA under the null hypothesis of no average causal effect ($H_0 : ACE_{10} = 0$). In line with the properties of the data generation procedure, we tested the robustness of various statistical methods against violations of their assumptions and addressed the following research questions and hypotheses:

1. The decomposition of the unit-covariate Z into the between-component Z_b and the within-component Z_w has to be accounted for in the statistical analysis of conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level, if Z_b and Z_w independently influence the conditional effect function $CCE_{jk;Z,Z_b}$. Neglecting the decomposition and specifying a naive adjustment model that uses only the covariate Z as predictor leads to a considerable bias in the *ACE*-estimator (Snijders & Bosker, 1999).

2. Even if the model is correctly specified in its fixed part and takes the decomposition of Z into the between-component Z_b and the within-component Z_w into account, residual effects of the cluster variable C need to be modeled by estimating additional variance components for $Var(r_{0;C})$ and $Var(r_{10;C})$ in the random part of the model. Statistical models that do not include these additional variance components will lead to standard errors that underestimate the variability of the *ACE*-estimates (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

3. Statistical models that include the appropriate variance components for $Var(r_{0;C})$ and $Var(r_{10;C})$, but do not treat the unit-covariate Z as a stochastic predictor and assume that the empirical mean of Z is equal to the expected value $E(Z)$ and constant over replications, will underestimate the variability of the corresponding *ACE*-estimates, especially when there are strong interactions between the treatment variable X and the covariates (Kröhne, 2009; Nagengast, 2006). Methods that take the stochasticity of the observed scores of Z explicitly into account by estimating $E(Z)$ as a model parameter will yield accurate standard errors of the *ACE*-estimator in all conditions.

Table 4.1: Properties of the statistical models to implement the generalized ANCOVA for designs with treatment assignment at the unit-level

Method	Decomposition of Z	Stochastic predictors	Variance components	Latent variable Z_b
`lace`: Naive model	-	x	-	-
`lace`: Full model	x	x	-	-
`nlme`: Full model	x	-	x	-
`Mplus`: Full model	x	x	x	-

Statistical details of the models and the implementation of the generalized ANCOVA for designs with treatment assignment at the unit-level are given in Appendix B. An overview of the properties of the statistical models with respect to the properties of the repeated single-unit trial and the resulting research questions is given in Table 4.1. Specifically, the following implementations of the generalized ANCOVA were studied:

- The *naive singlelevel model* implementation of the generalized ANCOVA in `lace` (Partchev, 2007), using only the unit-covariate Z as a predictor in separate group-specific structural equation models and neglecting the multilevel structure of the unit-covariate Z by not further decomposing Z into the between-component Z_b and the within-component Z_w [see Equation (B.5) in Appendix B.1 for the model specification];

- the *full singlelevel adjustment model* in `lace`, separately modeling the between-component Z_b and the within-component of Z_w, using the cluster-means of the unit-covariate Z, the cluster-mean centered values of Z and their product as predictors in group-specific structural equation models, but not including variance components for the residuals $r_{0;C}$ and $r_{10;C}$ [see Equation (B.6) in Appendix B.1 for the model speification];

- the *full multilevel adjustment model* in `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & the R Core team, 2008) separately modeling the between-component Z_b and the within-component of Z_w, using the cluster-means of the unit-covariate Z, the cluster-mean centered values of Z, the treatment indicator and their products as predictors, and modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by including a random intercept and a random effect of the treatment indicator,

but obtaining standard errors and significance tests of the *ACE* with the general linear hypothesis, thereby treating Z as a fixed predictor [see Equation (B.11) in Appendix B.2 for the model specification];

- the *full multilevel adjustment model* in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) specified as a singlegroup multilevel model, separately modeling the between-component Z_b and the within-component Z_w, using the cluster-means of the unit-covariate Z, the cluster-mean centered values of Z, the treatment indicator and their products as predictors, modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by including a random intercept and a random effect for the treatment indicator, and modeling the expected value of the unit-covariate Z as a model parameter, thereby taking the stochasticity of the unit-covariate Z explicitly into account in the estimation of the *ACE* and its standard error [see Equation (B.24) in Appendix B.3 for the model specification].

Unfortunately, it is not yet possible to specify the generalized ANCOVA for designs with treatment assignment at the unit-level as a multilevel latent variable model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) that includes the interaction of the dichotomous treatment indicator and the latent covariates Z_b and Z_w. The specification of a multigroup multilevel latent variable model in `Mplus` that could be naturally used to estimate these interactions requires that each cluster c only appears in one treatment group. For multilevel designs with treatment assignment at the unit-level, this requirement is violated: Each cluster can contain treated and untreated units and thus appear in both treatment group-specific models. Until now, this prohibits the application of the multigroup multilevel latent variable model to designs with treatment assignment at the unit-level (see also Appendix B.3) and had the unfortunate consequence that corrections for the unreliability of the empirical cluster means as measures for the between-component Z_b could not be tested in the simulation study.

There were several reasons, why the research questions with respect to the statistical methods required a simulation study and could not be addressed by analytical derivations only:

1. The distributional theory for all statistical methods considered holds only asymptotically (see Appendix B for further details) and the performance of these methods in finite samples under realistic circumstances determines their usefulness for applications.

2. The standard errors of the *ACE*-estimator in `lace` and `Mplus` obtained with the multivariate delta-method are — even asymptotically — first-order Taylor-series approximations (Rao, 1973; Raykov & Marcoulides, 2004). Thus, simulation studies are called for to analyze their properties and appropriateness in finite samples (see also MacKinnon, 2008).

3. Finally, none of the statistical methods correctly incorporated *all* peculiarities of the data generation (e.g., no method modeled Z_b and Z_w as latent variables or otherwise corrected for the unreliability of the cluster means of Z) — robustness against such violations determines the applicability of the statistical methods.

4.2.3 Design

The simulation was implemented in `R` 2.6.1 (R Development Core Team, 2008) using `SimRobot` (Kröhne, 2007) to manage and distribute the simulation conditions on a cluster of 40 workstations. All simulated data sets were generated with the data generation routine for designs with treatment assignment at the unit-level described in Appendix C.1. In the following section, the parameters that were varied — the independent variables of the simulation design — and the parameters that were held constant over simulation conditions are described in detail.

Independent Variables

The following parameters of the data generation were varied in a six-factorial fully-crossed simulation design with 1000 replications per cell: (1) the number of clusters, (2) the average number of units within each cluster, (3) the intraclass correlation coefficient of the unit-covariate Z, the partial regression coefficients of the logistic assignment function varied independently for (4) the within-component Z_w and (5) the between-component Z_b and (6) the effect size of the interaction between the treatment variable X and the between-component Z_b. We will describe the values of the design factors and outline the motivation behind these choices. An overview of the design factors is given in Table 4.2. The corresponding parameters of the data generation routine are given in Table C.1 in Appendix C.1.

Table 4.2: Factors of the simulation design for designs with treatment assignment at the unit-level

Factor	Measure	Values
Number of clusters		20, 50, 200
Average cluster sizes	$\overline{n}_c$	50, 100, 250
Intraclass correlation of Z	$ICC(Z)$	0.05, 0.1, 0.2, 0.3
Dependency of X and Z_w	$\sqrt{R^2_{inc}}$	0, 0.15, 0.3, 0.5
Dependency of X and Z_b	$\sqrt{R^2_{inc}}$	0, 0.15, 0.3, 0.5
Effect size of cluster-level interaction	$d(\gamma_{11})$	0, 0.1, 0.2, 0.3

Number of Clusters. The total number of clusters was either 20, 50 or 200. 20 clusters were chosen as a reasonable lower bound for the total number of clusters, since cost-effective designs for multisite randomized trials involve a relatively small number of clusters with a medium or large number of units per cluster (Moerbeek, van Breukelen, Berger, & Ausems, 2003; Raudenbush & Liu, 2000). The complexity of the adjustment model made it unlikely that the statistical models would perform well with a smaller number of clusters given the results of previous simulation studies: Acceptable performance of complex hierarchical linear models with random slopes has only been reported for samples as small as 50 clusters (Browne & Draper, 2000; Maas & Hox, 2005). 200 clusters were chosen as an upper bound for the number of clusters; larger samples are unlikely to be obtained in between-group multilevel designs with treatment assignment at the unit-level due to the high marginal costs involved in sampling additional clusters (Moerbeek, van Breukelen, Berger, & Ausems, 2003; Raudenbush & Liu, 2000). Unsatisfactory performance of adjustment methods at this number of clusters would render them unsuitable for applications.

Average Cluster Sizes. The average number of units per cluster $\overline{n}_c$ was chosen to be either 50, 100 or 250. The actual cluster sizes varied by 10% around the average size to represent naturally occurring variations in cluster sizes in designs with pre-existing clusters. However, the total sample size for each replication was fixed to the product of the number of clusters and the average cluster size. The average cluster sizes were chosen to represent a wide range of realistic values in applications. Since cost-effective

multisite designs usually use medium to large clusters and the marginal costs of adding another participant in a cluster is considerably smaller than adding another cluster (Moerbeek, van Breukelen, Berger, & Ausems, 2003; Raudenbush & Liu, 2000), an average cluster size of 50 was chosen as a reasonable lower bound and secured that both treatment groups would be adequately represented within each cluster, even under extreme treatment probabilities. Average cluster sizes of 100 were chosen to represent large, but still reasonable samples for applications and an average of 250 units per cluster was chosen as extreme and almost asymptotic value.

Intraclass Correlation of the Unit-Covariate. The following values of the intraclass correlations of the unit-covariate Z [$ICC(Z)$] were considered: $0.05, 0.1, 0.2, 0.3$. They were chosen to cover the range of typical intraclass correlation coefficients found in medical and educational research in accordance with the reviews by Gulliford et al. (1999), Hedges and Hedberg (2007) and Schochet (2008). An $ICC(Z)$ of 0.3 is higher than most of the empirical ICC-values reported in these studies. In educational studies, values between 0.1 and 0.2 are fairly common (Hedges & Hedberg, 2007; Schochet, 2008). In medical research, values of around 0.05 are normal, when reasonable cluster sizes are considered (Gulliford et al., 1999). An $ICC(Z)$ of zero, as expected in designs with random assignment of units to clusters, was excluded, since such designs are seldom implemented (Murray, 1998). The $ICC(Z)$ was manipulated by varying the variance of Z_b accordingly, while holding the variance of Z_w constant at a value of 1. The exact values of the corresponding variance parameters can be found in Table C.1 in Appendix C.1.

Dependency of Z_w / Z_b and X. The dependencies between the treatment variable X and the within-component Z_w [represented by the parameter g_1 in Equation (4.36)] and the treatment variable X and the between-component Z_b [represented by the parameter g_2 in Equation (4.36)] were varied independently to represent different degrees of confounding. For this purpose, the partial regression coefficients of the logistic assignment function given in Equation (4.36) — representing the independent influences of the between-component Z_b and the within-component Z_w on the treatment assignment probabilities — were chosen to the hold the square-root of the increment of the coefficient of determination of the logistic assignment function $\sqrt{R^2_{inc}}$ (Nagelkerke, 1991) constant at values of $0, 0.15, 0.3, 0.5$. When both parameters were set to zero, the treat-

ment assignment probabilities did not depend on the covariates and the design was actually a multisite randomized experiment with equal treatment probabilities in each cluster. The parameters of the logistic assignment function held the unconditional probabilities for treatment and control group equal in all conditions of the simulation design [$P(X{=}0){=}P(X{=}1){=}0.5$], no matter how strong the dependencies between the treatment variable X and the within-component Z_w and the between-component Z_b were. Since the variance of the between-component Z_b varied with different $ICC(Z)$-values, different values of g_2 had to be chosen to obtain constant values of $\sqrt{R^2_{inc}}$. The exact values of g_1 and g_2 were obtained in an exploratory simulation study and are exact up to the third decimal. The parameter values are given in Table C.1 in Appendix C.1.

Effect Size of Cluster-Level Interaction. Finally, the effect size of the interaction between X and Z_b was varied as a factor in the simulation design. In order to define an effect size measure for the interaction that was independent of the strength of association between X, Z_b and Z_w, the effect size was measured by the proportion of the sum of the variance of the true-outcome variable τ_1 in the treatment condition and the residual variance σ^2_Y of the outcome variable Y that was due to the interaction γ_{11}:

$$d(\gamma_{11}) = \frac{Var(\gamma_{11} \cdot Z_b)}{Var(\tau_1) + \sigma^2_Y} = \frac{\gamma^2_{11} \cdot Var(Z_b)}{Var(\tau_0 + \delta_{10}) + \sigma^2_Y}. \tag{4.37}$$

The regression weight γ_{11} is the regression weight of the product variable of X and Z_b in Equation (4.18). The effect size $d(\gamma_{11})$ was set to values of $0, 0.1, 0.2, 0.3$ to represent realistic to extreme effect sizes of interactions in applications. Since the variance of the between-component Z_b varied with different values of $ICC(Z)$, the corresponding parameter values for γ_{11} had to be varied accordingly to keep $d(\gamma_{11})$ constant for different conditions of $ICC(Z)$. These values were obtained analytically using `YACAS` (Goedman, Grothendieck, Højsgaard, & Pinkus, 2007). The corresponding parameters of the data generation function are documented in Table C.1. In all cases, the regression weight γ_{11} of the product of X and Z_b was positive or equal to zero, while the regression weight γ_{14} of the product of X and Z_w, that was not varied in the simulation, was negative.

Constant Parameters

Average Causal Effect. The goal of the simulation was to study the performance of the different implementations of the adjustment model under the null hypothesis of no average causal effect ($H_0 : ACE_{10} = 0$). The model parameters were chosen to guarantee that the average causal effect ACE_{10} was fixed at a value of zero in all experimental conditions. The expected value $E(Z)$ of the unit-covariate Z was set to 1. Since, the regression weight γ_{11} of the product of X and Z_b varied with the effect size of the interaction and different values of $ICC(Z)$, the intercept γ_{10} of the conditional effect function $CCE_{10;Z,Z_b}$ was varied accordingly to fix the average causal effect to zero in all cells of the simulation design. An overview of all fixed parameters is given in Table C.2 in Appendix C.1.

Variance Parameters. The following variance parameters were constant over all experimental conditions: $\sigma^2_Y = 2$, $\sigma^2_{Z_w} = 1$, $\sigma^2_{\upsilon_{0;U}} = 2.25$, $\sigma^2_{\upsilon_{10;U}} = 1.25$, $\sigma^2_{r_{0;C}} = 0.75$ and $\sigma^2_{r_{10;C}} = 0.25$. These settings resulted in a moderate heterogeneity of unit- and cluster-level variances of the outcome variable Y. However, the (X=j)-conditional residual intraclass correlation coefficients $rICC_{X=j}(Y \mid Z, Z_b)$ of the outcome variable Y were similar in treatment and control group (0.15 in the control group, 0.153 in the treatment group). The residual design effect (Kish, 1965) was larger than 2 in all sample size conditions (all residual VIFs > 2). Analyses that neglected the multilevel structure of the data would result in significantly underestimated standard errors of the model parameters (Hox, 2002; Maas & Hox, 2005). The average reliabilities of the cluster means as measures of the between-component Z_b ranged from 0.725 in the conditions with the smallest average cluster sizes and the smallest $ICC(Z)$ to 0.991 in the conditions with the largest average cluster sizes and the largest $ICC(Z)$. An overview of all fixed variance parameters of the simulation design is given in Table C.2 in Appendix C.1.

4.2.4 Results

In this section, we present the main results of the simulation study. We begin with reporting the convergence rates of the different methods, then give the bias of the ACE-estimators and their standard errors. Finally, we report the empirical type-1-error rates

of the significance tests and briefly compare the mean squared errors of selected implementations. The dependent measures are introduced in Appendix D together with the cut-off criteria considered as boundaries for appropriate performance. At the end of the section, we summarize and review the results with respect to address the research questions introduced above. We only present selected graphics of the results in this section. All results for the different methods and dependent measures are provided as graphics on the accompanying CD, as are the raw results for all simulation conditions (see Appendix E).

Convergence

There were substantive convergence problems for the implementation of the adjustment model in `Mplus`. The convergence problems were independent of the $ICC(Z)$ and the effect size of the between-cluster interaction $d(\gamma_{11})$, but depended on the number of clusters, the amount of between-cluster and within-cluster confounding and the average cluster sizes. Most strikingly, the `Mplus` model implementation had strong convergence problems when the between-component Z_b did not influence the treatment assignment probabilities (an average convergence rate of 61.78% in these cells versus an average convergence rate of 93.64% in all other conditions). If the between-component Z_b influenced treatment assignment, non-convergence was only problematic for the smallest number of clusters: In the conditions with 20 clusters, an average of 81.30% of the replications did converge, whereas 99.62% of the models did converge for 50 clusters and all models converged for 200 clusters. In the conditions in which the between-component Z_b did not influence the treatment assignment probabilities, the convergence rates were further moderated both by the average cluster size and by the dependency of the within-component Z_w and the treatment variable X. Convergence rates were smaller when the within-component did not influence the treatment assignment and increased with increasing influence of Z_w. The worst convergence rates were obtained in cells, where neither the between-component Z_b nor the within-component Z_w influenced the treatment assignment, i.e., in multisite randomized trials with equal treatment probabilities for all clusters and units. Surprisingly, this effect was further augmented with larger average cluster sizes. Additional exploratory simulations showed that the convergence problems were neither alleviated by using the true parameter values as starting values nor by further increasing the number of iterations or loosening the convergence criteria.

The convergence pattern for the `Mplus` model implementation is given in Figure 4.1 on the following page.

The implementations in `nlme` and `lace` did not show any convergence problems. All replications in all experimental cells converged.

Bias of *ACE*-Estimator

In this section, we describe the bias in estimation of the average causal effect for the different implementations of the generalized ANCOVA. Our discussion will be organized as follows: We start with presenting the results for the naive adjustment model in `lace`, followed by a discussion of the full adjustment models that used the empirical cluster-means and cluster-mean centered values of the unit-covariate Z as predictors to take the multilevel structure of the effects of the unit-covariate Z into account. We use the mean bias (*MB*) of the *ACE*-estimator as defined in Equation (D.1) to evaluate the different estimators. Following the recommendations by Boomsma and Hoogland (2001), an over- or underestimation of the *ACE* by 2.5% (corresponding to a *MB* between -0.025 and 0.025) was considered as threshold for unbiasedness of an estimator.

Naive Adjustment Model in `lace`. The naive adjustment model implementation in `lace` that only included the unit-covariate Z and did not account for differential influences of the within-component Z_w and the between-component Z_b overestimated the average causal effect (average mean bias over all conditions: $MB = 0.074$). A closer inspection of the results revealed that the *MB*s varied considerably between the conditions of the simulation designs: The absolute value of the *MB* was over the threshold of 0.025 in a total of 1882 cells of the simulation design, indicating that 81.68% of all cells exhibited a significant bias. The *ACE* was overestimated in 1146 cells (49.74% of all cells) and underestimated in 736 cells (31.94% of all cells). As these numbers indicate, the naive model in `lace` is not a suitable implementation of the generalized ANCOVA and a detailed discussion of the pattern of the bias patterns is foregone.

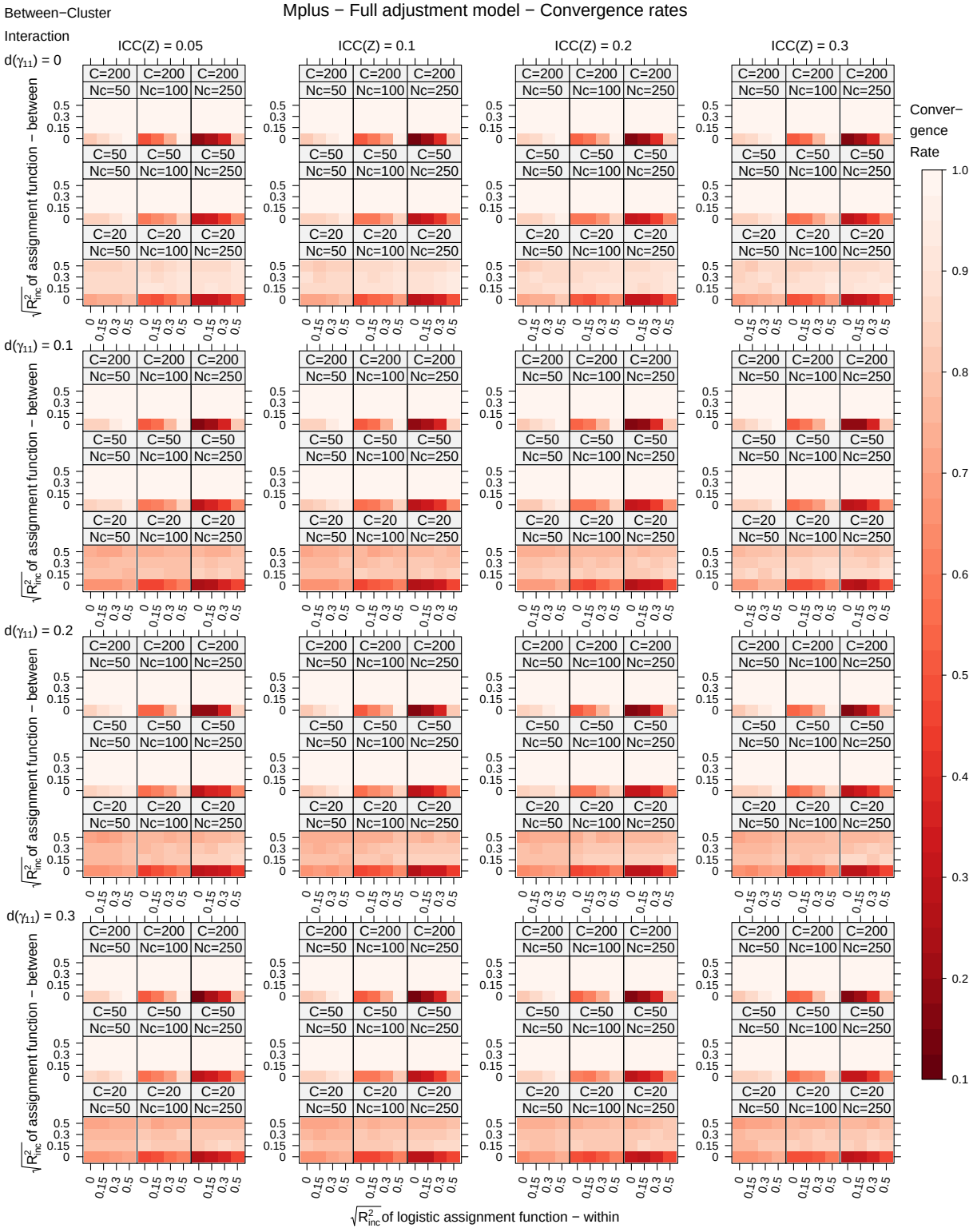

Figure 4.1: Convergence rates: Full adjustment model implemented as singlegroup multilevel model in `Mplus`

Full Adjustment Model in `lace`. The full adjustment model as implemented in `lace` included both the between- and the within-component of the unit-covariate (modeled with the empirical cluster means and the cluster-mean centered raw scores of the unit-covariate), but was restricted to a singlelevel model and did not include the additional variance components for the intercept and the random slope of the treatment variable. The average mean bias of the *ACE*-estimator over all conditions was $MB = 0.010$. Although on average, the model performed well, a total of 464 (equal to 20.14%) cells had an absolute bias above the threshold of 0.025. 91 cells (3.95%) yielded a negatively biased *ACE*-estimator, 373 cells (16.19%) yielded a positively biased *ACE*-estimator. While the model performed well when no interaction at the between-cluster level was present, overestimation became more critical with larger effect sizes of the interaction at the between-level. Overestimation of the *ACE* was especially prevalent in cells with small $ICC(Z)$-values and a strong dependency between the between-component of Z_b and the treatment variable X. This effect was more pronounced with larger interaction effects on the cluster-level. The full pattern of results is shown in Figure 4.2 on the next page.

Full Adjustment Model in `nlme`. The full adjustment model in `nlme` included the between-component Z_b and the within-component Z_w (modeled with the empirical cluster means and cluster-mean centered values of the unit-covariate) as well as the variance components for the intercept and the random slope. The average mean bias of the *ACE*-estimator was very close to zero $MB < 0.001$, indicating an almost perfect estimation of the average causal effect on average. Further inspection of the results revealed that only 37 cells (1.61% of all cells) had an absolute *MB* above the critical value of 0.025, also indicating unbiasedness of the *ACE*-estimator.

Full Adjustment Model in `Mplus`. The full adjustment model in `Mplus` also included the between-component Z_b and the within-component Z_w (modeled with the empirical cluster means and cluster-mean centered values of the covariate) as well as the variance components for intercepts and random slopes. The average mean bias of the *ACE*-estimator of this model implementation was also very close to zero $MB < 0.001$. Further inspection of the results revealed that only 69 cells (corresponding to 2.91% of all cells) had an absolute *MB* above the threshold of 0.025, indicating unbiasedness of the *ACE*-estimator for almost all conditions.

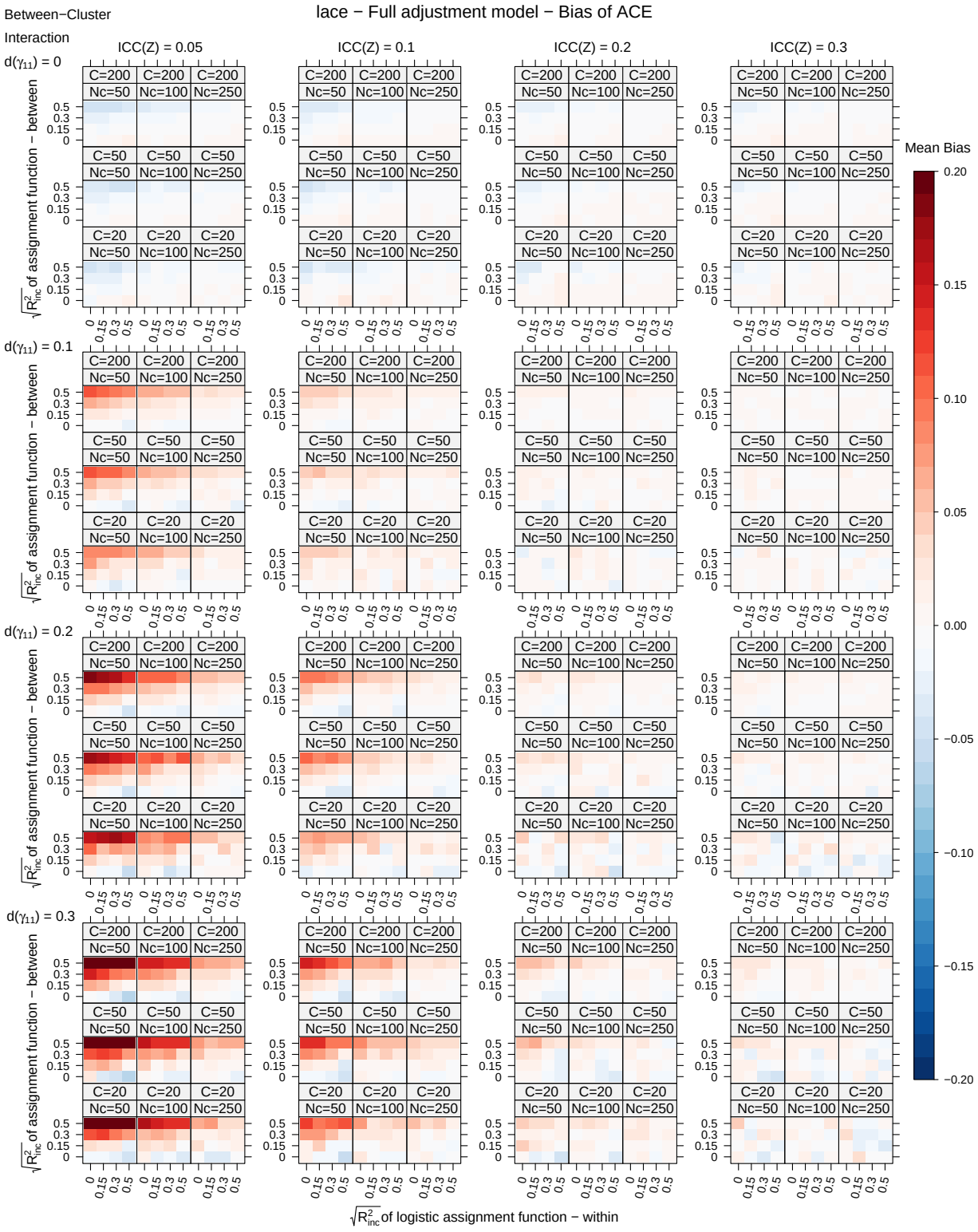

Figure 4.2: Mean bias of *ACE*-estimator: Full adjustment model in `lace`

Bias of Standard Error

In this section, we discuss the bias of the standard error of the *ACE*-estimator focusing on the models that yielded an unbiased or relatively little biased *ACE*-estimator. Since the naive model implementation in `lace` resulted in a strongly biased *ACE*-estimator, we will not discuss the standard errors of this implementation in more detail. We use the mean relative bias (*MRB*) as defined in Equation (D.3) to evaluate the standard error estimators. Following the recommendations by Boomsma and Hoogland (2001), the standard error estimator was considered unbiased, if it over- or underestimated the empirical variability of the *ACE*-estimator by less than 5% (corresponding to a *MRB* between −0.05 and 0.05). We start with presenting the results of the standard error estimator in `lace`, followed by the standard error for the *ACE*-estimator in `nlme` and in `Mplus`.

Full Adjustment Model in `lace`. The average *MRB* of the standard error of the *ACE*-estimator in the implementation of the full adjustment model in `lace` was −0.599, indicating a strong negative bias of the standard error of the *ACE*-estimator. The *MRB* of the standard error was below the threshold of −0.05 in all cells of the simulation designs, indicating that the standard error significantly underestimated the empirical variability of the *ACE*-estimator in all conditions.

Full Adjustment Model in `nlme`. The average *MRB* of the standard error of the *ACE*-estimator in the implementation of the full adjustment model in `nlme` was *MRB* = −0.390, indicating a strong negative bias of the standard error of the *ACE*-estimator. This bias was due to the conditions in which the effect size of the interaction $d(\gamma_{11})$ was larger than zero: Without an interaction between X and Z_b, the average *MRB* was close to zero (average *MRB*=−0.002). In all other conditions with an interaction at the between-level, the *MRB* indicated a significant underestimation of the empirical variability of the *ACE*-estimator by the corresponding standard error (average *MRB*=−0.519). While all cells had an absolute *MRB*-value over the cut-off of 0.05 when interactions were present at the cluster-level, only 15 cells (corresponding to 2.60% of all remaining cells) had an absolute *MRB* over 0.05 when no interaction at the between-level was present.

Full Adjustment Model in `Mplus`. The average *MRB* of the standard error of the *ACE*-estimator in the implementation of the full adjustment model in `Mplus` was 0.037, indicating a small positive bias of the standard error estimator over all conditions of the simulation design. This bias was mostly driven by the conditions with the smallest number of clusters and no interactions at the between-level: In these cells, the average *MRB* of the standard error estimator was $MRB = 0.391$, indicating that the standard error overestimated the empirical variability of the *ACE*-estimator considerably. When these conditions, that also suffered from severe convergence problems, were excluded from the analysis, the average *MRB* dropped to 0.005. Of the remaining simulation cells 168 (or 7.95%) had a *MRB* above the upper threshold of ($MRB > 0.05$) and 54 cells (or 2.56%) had a *MRB* below the lower threshold ($MRB < -0.05$). The larger number of cells with critical overestimation of the standard error was due to a positive bias in conditions with a small effect size of the interaction, large $ICC(Z)$ values and the smallest number of clusters considered. No other combination of simulation conditions influenced the bias of the standard error in a systematic way. An overview of the bias pattern for the standard error is given in Figure 4.3 on the following page.

Type-1-Error Rates

In this section, we present the results of the empirical type-1-error rates for two-sided significance tests of the null hypothesis of no average causal effect at an α-level of 0.05. Tests at different significance levels yielded similar results. In line with the suggestions by Boomsma and Hoogland (2001), the limits of the 95%-confidence interval for the rejection frequency of an adequate significance test were calculated (for details see Appendix D) and used to evaluate the performance of the significance tests. The lower limit of the confidence interval for 1000 replications was 0.037; the upper limit was 0.064. We will only report the results for the implementations of the full adjustment model.

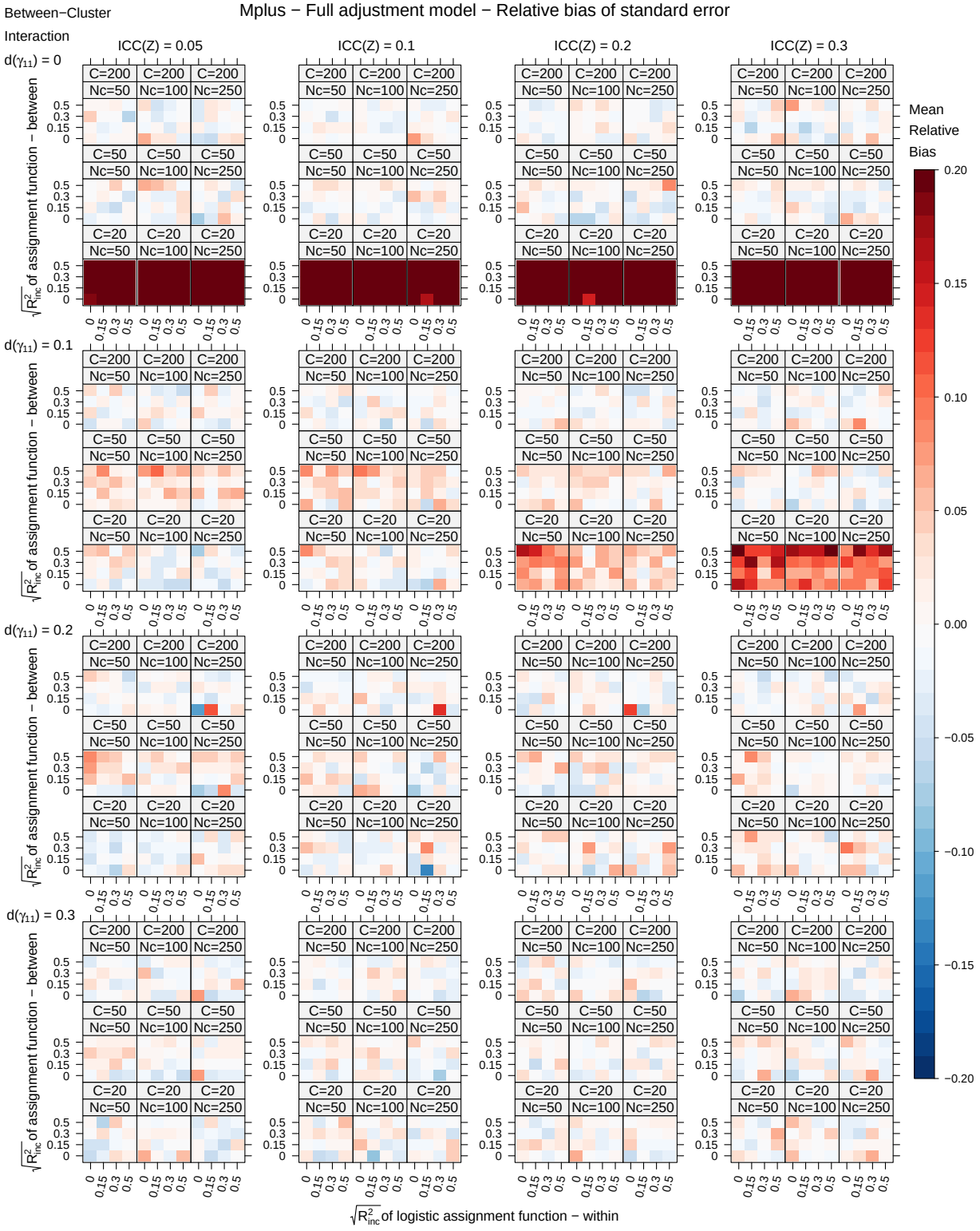

Figure 4.3: Mean relative bias of standard error: Full adjustment model implemented as singlegroup multilevel model in `Mplus`

Full Adjustment Model in `lace`. The mean type-1-error rate of the full adjustment model in `lace` averaged over all conditions of the experimental design was equal to 0.471. Even in the cell with the smallest type-1-error rate (type-1-error rate equal to 0.085), the nominal α-level was clearly exceeded. In line with the results of the standard error estimator, these findings indicated that significance tests of the null hypothesis of no average causal effect were clearly too liberal when the multilevel structure of the residual variances was not modeled properly.

Full Adjustment Model in `nlme`. The mean type-1-error rate of significance tests of the average causal effect in the implementation of the full adjustment model in `nlme` was 0.285. The significance test performed adequately when no interactions between Z_b and X were present (average empirical type-1-error rate equal to 0.055), but was clearly too liberal in conditions with interactions on the between-level (average type-1-error rate equal to 0.362). However, even in the former condition a total of 84 cells (corresponding to 14.58%) had an empirical type-1-error rate larger than 0.064 and thus fell outside of the 95%-confidence interval. In the conditions in which interactions on the between-cluster level were present, all cells had an empirical type-1-error rate larger than 0.064.

Full Adjustment Model in `Mplus`. The average type-1-error rate for the adjustment model implementation in `Mplus` was 0.059, indicating a slightly elevated empirical significance level. This result was mainly driven by heightened type-1-error rates in the condition with the smallest number of clusters (average type-1-error rate equal to 0.068) compared to the conditions with more number of clusters (average type-1-error rate equal to 0.055). In the latter conditions a total of 193 cells (corresponding to 12.57% of the remaining cells) had an empirical type-1-error rate larger than 0.064 and thus outside of the 95%-confidence interval, indicating that the significance test was slightly too progressive. When all cells with significant convergence problems were dropped from the analysis the average type-1-error rate dropped further to 0.053 and only 28 cells (or 4.62% of all remaining cells) had an empirical type-1-error rate above the upper limit of the 95%-confidence interval, indicating a satisfactory performance of the significance test.

Efficiency of *ACE*-Estimator

In order to study the efficiency of the unbiased *ACE*-estimators, we compared the mean squared errors (*MSE*) of the implementation of the adjustment model in `nlme` to the *MSE* of the implementation of the adjustment model in `Mplus` by computing the ratios of the *MSE*s in each cell of the simulation design. Values larger than 1 indicated cells in which the `Mplus` implementation was more efficient, values smaller than 1 indicated cells in which the `nlme` implementation was more efficient. The mean ratio over all cells of the simulation design was 0.965, indicating that the `nlme` implementation was on average slightly more efficient than the `Mplus` model. This result was mainly driven by the conditions with the smallest number of clusters (mean *MSE*-ratio=0.885) in which the `Mplus` model had convergence problems and `nlme` clearly outperformed `Mplus` in efficiency. In the conditions with more than 20 clusters the average *MSE* ratio was equal to 1.006 and indicated that both implementations of the adjustment model were equally efficient on average. In 808 cells (corresponding to 52.60% of the remaining conditions), the `Mplus` model was more efficient, in the remaining 728 cells (corresponding to 47.40% of the remaining conditions) the `nlme` model was more efficient though no distinctive pattern of differences emerged.

Summary of Results

The results can be summarized with respect to the research questions introduced in Section 4.2.2 as follows:

1. The simulation study showed that the decomposition of the unit-covariate Z into the between-component Z_b and the within-component Z_w and their differential effects on the outcome variable Y had to be explicitly modeled in the generalized ANCOVA for multilevel designs with treatment assignment at the unit-level. The naive model implementation in `lace` that only included the unit-covariate Z yielded a biased *ACE*-estimator. The direction of the bias depended on the specific within- and between-cluster effects and interactions, so that no general correction formula can be given without explicitly accounting for the multilevel structure of the design.

2. The singelevel implementation of the full adjustment model in `lace` that correctly modeled the within- and the between-component of the unit-covariate, but did

not include the variance components for the residuals $Var(r_{0;C})$ and $Var(r_{10;C})$ did not yield correct standard error estimators. The empirical variability of the corresponding *ACE*-estimator was underestimated in almost all conditions of the simulation design. Additionally, the *ACE*-estimator was biased in some conditions.

3. Finally, we showed that the implementation of the adjustment model in `nlme` that included a random intercept and a random slope for the treatment indicator yielded an unbiased *ACE*-estimator, but underestimated its variance considerably by not modeling stochastic predictors. On the other hand, the implementation of the adjustment model in the multilevel structural equation model in `Mplus` that included the estimator of the expected value of Z as a model parameter and calculated the standard error with the multivariate delta method yielded accurate standard errors.

In the next section, we will apply the different implementations of the generalized ANCOVA to an illustrative example. We will further discuss the results of the simulation study in Section 4.4.

4.3 Example Analysis

In this section, we illustrate the statistical implementations of the generalized ANCOVA for designs with treatment assignment at the unit-level with an empirical example from the National Educational Longitudinal Study of 1988 (NELS:1988, Curtin et al., 2002). We estimate the average effect of participating in academic clubs (such as biology or mathematics clubs) on a science test for high school sophomores (corresponding to grade ten in the German educational system) controlling for the between- and within-effects of the science pre-test obtained two years earlier during middle school (corresponding to grade eight). The analyses are intended to demonstrate the importance of adjusting average treatment effects for the influences of covariates in observational studies and quasi-experiments and the flexibility of the generalized ANCOVA in doing so; they are not aimed at deriving substantive insights into the effects of participating in academic clubs on high school achievement. Conceptually, the data comes from a quasi-experiment with treatment allocation at the unit-level, with non-random

assignment of units to clusters and self-selection to treatment conditions. The causal interpretation of the obtained average effects rests on the assumption of an unbiased unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, Z_b)$ — an assumption that is not tested explicitly in the analyses.

4.3.1 Methods

Design

The National Educational Longitudinal Study of 1988 (NELS:1988) is a longitudinal study that followed US high school students in the 1990s. The description of the design and the materials used in NELS:1988 follows the user manual for the public-use data (Curtin et al., 2002). The initial data collection took place in 1988, when the studied cohort was still in middle school, and was continued with two follow-ups when the students were high school sophomores (1990) and shortly after they had finished high school (1992). A fourth follow-up was conducted in 2000 to assess postsecondary educational outcomes and the transition to the labor market. Here, we are using data from the base year assessment and the first follow-up. We study the effects of participating in academic clubs on a science test controlling for student and school characteristics on a sample of 9112 students in 1281 schools with complete data for all variables. The average cluster size was 7.113. The smallest school sample consisted of only one student, the largest school sample contained 38 students.

NELS:1998 used a complex sample design to obtain a nationally representative sample of students (see, Curtin et al., 2002). Analyzing the data with the goal of obtaining nationally representative results would require a complex weighting scheme to account for design effects, non-response and dropout. Since we only use this data to illustrate different implementations of the generalized ANCOVA, we ignore this additional complexity and analyze the data as if it was obtained from a simple random sample.

Materials

The science test contained 25 questions from life sciences, earth sciences, physical sciences, chemistry and scientific method that were answered as part of a larger cognitive assessment battery of 116 questions in a total of four subject areas (reading comprehension, mathematics, science and social studies) to be completed in 85 minutes. The

items of the science test were designed to assess scientific knowledge, scientific understanding and problem solving. Individual test scores for the pre-test and post-test on a common scale were obtained with IRT scaling (details are given in Rock, Pollack, & Quinn, 1995). The IRT-estimator of the science ability was scaled in the T-metric with a mean of 50 and a standard deviation of 10. The average reliability of the IRT scores for the science test — computed as one minus the ratio of the average measurement error variance to the total variance (Rock et al., 1995) — was 0.73 at the base-year assessment and 0.81 at the first follow-up.

Information about the treatment variable — participation in academic clubs — was obtained from a self-administered questionnaire filled out by the students at the first follow-up. Overall 3124 students indicated that they had participated in an academic club, 5988 students indicated that they had not participated in an academic club during the first two high school years. The school identification number at the first follow-up served as cluster variable. The empirical cluster means of the science pre-test score were computed by averaging over the pre-test values of the students within a school. Since not all students within a school were sampled, the empirical cluster means were fallible measures of the latent cluster-covariate Z_b (see also Lüdtke et al., 2008).

Statistical Procedures

We compared the following implementations of the generalized ANCOVA for designs with treatment assignment at the unit-level. The models are the same that were studied in the simulation study. Specifically, we estimated the *ACE* with

(a) the naive model in `lace`,

(b) the full adjustment model in `lace`,

(c) the full adjustment model in `nlme`,

(d) the full adjustment model in `Mplus`.

The full description of the adjustment models are given in Appendix B. Additionally and for comparison, we also computed the unadjusted treatment effect in `nlme`. As suggested by Steyer and Partchev (2008), effect sizes $d(\widehat{ACE})$ for the average effect estimates were obtained by dividing the *ACE*-estimate with the standard deviation of the outcome variable in the control group. Estimates of the intraclass correla-

Table 4.3: Descriptive statistics of the variables in the NELS:1988 data set

Variable	Mean	*SD*	*ICC*	Min	Max
Science Pre-Test	46.256	8.268	0.168	22.51	69.42
Science Post-Test	50.956	10.037	0.206	25.43	74.92
Treatment	0.343	0.475	0.266	0	1

tion coefficients *ICC* of the pre- and post-test measures were obtained from intercept-only models specified in `nlme`. The *ICC* of the treatment variable was estimated with an intercept-only model specified as a logistic linear mixed model using the `R`-package `MASS` package (Venables & Ripley, 2002) following Goldstein, Browne, and Rasbash (2002, Method C). The influence of the between-component Z_b and the within-component Z_w on the treatment probabilities at the between- and the within-level were estimated obtained with a logistic mixed regression model in `MASS`.

4.3.2 Results

Intercept-Only Models. The intercept-only model for the pre-test of scientific knowledge indicated that a significant amount of variance of the pre-test was located between schools, i.e., that schools differed in their expected values of the pre-test scores ($\hat{\sigma}^2_{Z_b} = 11.556; \hat{\sigma}^2_{Z_w} = 57.213; \widehat{ICC}(Z) = 0.168$). The estimated intercept parameter was $\hat{\gamma}_{00} = 46.124$ ($SE = 0.134, t = 344.209, p < 0.001$). The average reliability of the cluster-means as indicators for the latent between-component Z_b was 0.590 as calculated according to Equation (2.10).

The intercept-only model for the post-test of scientific knowledge also indicated that a significant amount of variance was located between schools, i.e., that schools differed on the average value of the science post-test ($\hat{\sigma}^2_{Z_b} = 20.927; \hat{\sigma}^2_{Z_w} = 80.567; \widehat{ICC}(Z) = 0.206$). The estimated intercept parameter was $\hat{\gamma}_{00} = 50.840$ ($SE = 0.172, t = 295.581$, $p < 0.001$), roughly four-and-a-half points higher than for the pre-test.

The intercept-only model for the treatment variable, participation in academic clubs, showed that a significant amount variance was located between schools, i.e., that schools differed in the average treatment probabilities ($\hat{\sigma}^2_{X_b} = 0.340; \hat{\sigma}^2_{X_w} = 0.937; \widehat{ICC}(Z) = 0.266$).

Dependency of Covariates and Treatment. The logistic regression of the treatment variable on the cluster-means of the science pre-test and the cluster-mean centered individual scores of the science pre-test indicated that they both influenced treatment assignment probabilities significantly ($\beta_w = 0.038, SE = 0.003, t = 12.086, p < 0.001$; $\beta_b = 0.035, SE = 0.006, Z = 5.582, p < 0.001$). The corresponding values of Nagelkerke's (1991) $\sqrt{R^2_{inc}}$ were 0.084 for the cluster means of the pre-test and 0.141 for the cluster-mean centered individual pre-test scores. These results indicated that the probability of attending an academic club became larger with the average pre-test value in a school and for students with higher individual pre-test values relative to the average value in their school.

Unadjusted Treatment Effect. The unadjusted treatment effect obtained with `nlme` was $\hat{\gamma}_{10} = 2.702$ ($SE = 0.225, t = 11.988, p < 0.001$). This result indicated a medium positive effect of participating in academic clubs on science knowledge [$d(\widehat{ACE}) = 0.272$], if no other covariates were considered. The supposed treatment effect was statistically significant at a two-tailed significance-level of 0.05. For comparison purposes, the treatment effects of all implementations are displayed in Table 4.4 on the next page.

Naive Adjustment Model in `lace`. The naive adjustment model in `lace` that included only the raw scores of the pre-test as predictors and did not take the multilevel structure of the data into account, estimated the average causal effect as $\widehat{ACE} = -0.124$ ($SE = 0.077, t = -1.602, p = 0.110$). It suggested a small negative treatment effect [$d(\widehat{ACE}) = -0.012$] that was not statistically significant at a two-tailed significance-level of 0.05.

Full Adjustment Model in `lace`. The full adjustment model in `lace` that included the cluster-means of the pre-test, the cluster-mean centered individual scores of the pre-test and their interactions as predictors in a two-group structural equation model, but no additional variance components, estimated the average causal effect of the treatment with a non-linear constraint of the model parameters as $\widehat{ACE} = 0.737$ ($SE = 0.151, t = 4.893, p < 0.001$). This result indicated a small positive average effect of the treatment [$d(\widehat{ACE}) = 0.074$] that was statistically significant at a two-tailed significance-level of 0.05.

Table 4.4: Comparison of the estimated *ACE*s of the different adjustment procedures for the NELS:1988 data set

Method	$\widehat{ACE}$	$d(\widehat{ACE})$	SE	t-value	p-value	95%-conf.-interval
`nlme`: No adjustment	2.702	0.272	0.225	11.988	< 0.001	[2.260; 3.144]
`lace`: Naive model	−0.124	−0.012	0.077	−1.602	0.110	[−0.276; 0.028]
`lace`: Full model	0.737	0.074	0.151	4.893	< 0.001	[0.442, 1.032]
`nlme`: Full model	0.752	0.076	0.154	4.872	< 0.001	[0.450; 1.055]
`Mplus`: Full model	0.738	0.074	0.160	4.618	< 0.001	[0.424; 1.052]

Full Adjustment Model in `nlme`. The implementation of the full adjustment model in `nlme` included the cluster-means of the science pre-test, the cluster-mean centered individual scores of the science pre-test, the treatment variable and their second- and third-order interactions as predictors. As in the simulation study, a random intercept and a random slope for the treatment indicator were included. A specification test indicated that including a random effect of the within-component Z_w resulted in a better model fit ($\chi^2 = 22.908, df = 1, p < 0.001$), thus this random effect was additionally included. The *ACE*-estimate was not influenced by this inclusion. The average causal effect of the treatment estimated by this model with the general linear hypothesis was $\widehat{ACE} = 0.752$ ($SE = 0.154, z = 4.872, p < 0.001$). The full adjustment model with all interactions between the predictors in `nlme` thus suggested a small positive average treatment effect [$d(\widehat{ACE}) = 0.076$] that was statistically significant at a two-tailed significance-level of 0.05. The residual intraclass correlation coefficient *rICC* of the outcome variable after accounting for all predictors was 0.087. All model parameters of the full adjustment model in `nlme` are given in Table 4.5.

Full Adjustment Model in `Mplus`. The implementation of the full adjustment model in `Mplus` included the cluster-means of the science pre-test, the cluster-mean centered individual scores on the science pre-test, the treatment variable and their second- and third-order interactions as predictors. Additionally, a random intercept and a random slope for the treatment indicator were included. All unit-level predictors were allowed to have random slopes. The average causal effect of the treatment estimated by this model as a non-linear constraint of the model parameters was $\widehat{ACE} = 0.738$ ($SE = 0.160, z = 4.618, p < 0.001$). The full adjustment model with all interactions between the predictors in `Mplus` thus suggested a small positive average treatment effect [$d(\widehat{ACE}) = 0.074$] that was statistically significant at a two-tailed significance-level of 0.05. The standard error was slightly larger than the standard error obtained from `nlme`. Both implementations, however, would lead to the same conclusions about the average causal effect of participating in academic clubs on the science test.

4.3.3 Discussion

The results of the illustrative analysis mirrored the findings of the simulation study in many ways. Although the sample sizes at the cluster- and the unit-level were not ex-

Table 4.5: Parameters of the full adjustment model in `nlme`

Parameter	*Estimate*	*SE*	*t*-value	95%-conf.-interval
Fixed Effects				
γ_{00}: Intercept	3.531	1.054	3.351	[1.465; 5.597]
γ_{01}: Z_b	1.020	0.023	44.741	[0.975; 1.065]
γ_{04}: Z_w	0.517	0.153	3.388	[0.217; 0.817]
γ_{05}: $Z_w \cdot Z_b$	0.006	0.003	1.835	[0.000; 0.012]
γ_{10}: X	−1.991	1.653	−1.205	[−5.230; 1.245]
γ_{11}: $X \cdot Z_b$	0.059	0.035	1.674	[−0.010; 0.128]
γ_{14}: $X \cdot Z_w$	−0.064	0.258	−0.247	[−0.570; 0.442]
γ_{15}: $X \cdot Z_w \cdot Z_b$	0.002	0.006	0.350	[−0.010; 0.014]
Variance Components				
Level 1				
σ^2_ε	41.412			
Level 2				
$\sigma^2_{u_0}$	3.934			
$\sigma^2_{u_1}$	0.658			
$\sigma^2_{u_2}$	0.018			

plicitly included in the simulation — the average cluster size was considerably smaller than any condition of the simulation study and the cluster sizes varied more strongly; on the other hand the number of clusters was considerably larger — the other properties of the data were represented by the factors in the simulation study: The *ICC* of the science pre-test was equal to 0.168 and thus between two conditions of the simulation study. The $\sqrt{R^2_{inc}}$ of the dependencies between the treatment variable X and the between- and within-component of the pre-test were at the lower end of the simulated conditions and the effect size of the cluster-level interaction $d(\gamma_{11})$ was close to zero.

In line with the results of the simulations study, the *ACE*-estimates obtained from the implementations of the full adjustment model implementations in `lace`, `nlme` and `Mplus` did not differ considerably. They all indicated a small positive average effect of participating in voluntary academic clubs on scientific knowledge after accounting for the within- and between-component of the pre-test. The standard errors also behaved as expected: Of the three implementations, the estimate in `lace` had the smallest standard error, followed by `nlme` and `Mplus`. Due to the small effect size of the cluster-level

interaction and the relatively small residual intraclass correlation coefficient (*rICC*) of the outcome variable, the differences between the three standard errors were small and the confidence intervals around the *ACE*-estimates were very similar.

Nevertheless, the analyses demonstrated the importance of taking the between- and the within-component of the pre-test explicitly into account in the specification of generalized ANCOVA: The implementation of the naive adjustment model in `lace` that only included the raw scores of the pre-test as predictors would lead to vastly different conclusions about the average effect of participating in voluntary academic clubs on competence in science: In contrast to the appropriately specified models that accounted for the differential effects of the pre-test on the unit- and the cluster-level, the resulting *ACE*-estimate did not differ significantly from zero. The unadjusted treatment effect would have been similarly misleading: Without adjusting for the between- and the within-component of the pre-test, the average effect of attending an academic club would have been estimated to be larger than after controlling for these variables.

4.4 Discussion

In this section, we will discuss the adjustment model for multilevel designs with treatment assignment at the unit-level, its implementation in statistical models and the results of the simulation study and the example analysis. First, we review the problems of the promising statistical methods in the simulation study. Next, we discuss the limitations of the simulation study and outline further research needs. Then, we revisit the empirical example and its relation to the simulation study. Finally, we return to designs with unbiasedness of $E(Y \mid X, Z, C)$ that were not considered in a simulation and discuss several options to implement the generalized ANCOVA for these designs. We conclude the section with some recommendations for the application of the generalized ANCOVA for multilevel designs with treatment assignment at the unit-level. A comprehensive discussion of the merits and shortcomings of the generalized ANCOVA and the distinction between stochastic and fixed covariates will be given in the general discussion in Chapter 6.

4.4.1 Problems of the Statistical Models

Overall the results of the simulation study did not clearly favor one implementation of the full adjustment model for practical purposes. On the contrary, the two methods that yielded an unbiased *ACE*-estimator also had significant problems: The implementation of the adjustment model in a conventional linear mixed model using `nlme` (Pinheiro et al., 2008) converged reliably, but heavily underestimated the standard error in the presence of interactions between the treatment variable X and the between-component Z_b. The implementation in a singlegroup multilevel structural equation model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) yielded correct standard errors, but had severe convergence problems in many conditions of the simulation design. The implementation of the full and the naive adjustment model in `lace` resulted in biased *ACE*-estimators and underestimated the variability of the *ACE*-estimator, as expected.

The convergence problems of the `Mplus` implementation were caused (1) by an insufficient sample size relative to the number of parameters included in the model (this accounted for the convergence problems at conditions with 20 clusters) and (2) by an overparametrization of the generalized ANCOVA model in conditions where the between-component Z_b did not influence the treatment assignment probabilities. Surprisingly, the `nlme`-implementation, although specifying a similarly complex model did not show this convergence problems which points to instabilities of the implementation of the estimation algorithm in `Mplus` (see also, Pinheiro & Bates, 2000; Bates, Maechler, & Dai, 2008, for discussions of robust estimation of linear mixed effect models). The convergence problems of `Mplus` were most pronounced for multisite randomized trials, when neither the between-component Z_b nor the within-component Z_w influenced the treatment assignment probabilities. In these conditions, a simple mean comparison of the treatment and control group accounting for the effects of clustering (Hedges, 2007a; Raudenbush & Liu, 2000) would have been sufficient for obtaining an unbiased estimate and a correct standard error of the *ACE*. To simplify the model and make estimation more stable in applications, it seems advisable to sequentially test the necessity of including additional interactions and variance components. The convergence problems of the `Mplus` in cells in which the between-component Z_b did not influence the treatment probabilities, highlight the importance of selecting only covariates that influence both the outcome variable and treatment assignment at the same time to arrive at a parsimonious and stable adjustment model. Although no method

could be unequivocally recommended — especially for the smallest number of clusters — the adjustment model implementation in `Mplus` performed considerably better for larger number of clusters and larger cluster sizes. While it is rather unlikely that conditionally randomized designs are implemented in such huge samples, data from quasi-experimental evaluations might in some cases fulfill these sample requirements especially for designs with self-selection to conditions, but also with other-selection (e.g., when archival records are used to evaluate the nationwide effects of a medical treatment in multiple clusters, see Turpin & Sinacore, 1991). The implementation of the generalized ANCOVA in `nlme`, on the other hand, exhibited such a strong bias in the presence of interactions between X and Z_b, that it cannot be recommended at all, if such interactions are present.

4.4.2 Limitations of the Simulation Study

As in all simulations, the results of the present study cannot be generalized over and above the conditions realized in the data generation (Skrondal, 2000). The independent variables of the simulation design covered a wide range of realistic parameter values that were representative of non-randomized designs with treatment assignment at the unit-level that usually employ a small number of clusters with moderate to larger number of units per cluster (Turpin & Sinacore, 1991). Nevertheless, there were some notable structural limitations to the design of the study. In the following section, we first discuss the performance of the inappropriate statistical methods and review the properties of the data generation procedure that put them at a disadvantage. We then briefly discuss the omission of the conventional multilevel ANCOVA without interactions from the simulation design. Next, we reconsider variance heterogeneity that did not influence the results of the simulation, although it has been shown to be influential in previous studies (Kröhne, 2009). Finally, we outline further research needs to evaluate the performance of the statistical models.

Inappropriate Statistical Models

The data generation procedure was explicitly modeled after the multilevel single-unit trial introduced in Chapter 2 and consisted of a series of repetitions of independent and stable single-unit trials. This stacked the deck against the conventional hierarchical linear model in `nlme` and the singlelevel implementation of the generalized ANCOVA

in `lace` in four ways, that should be taken into account when trying to generalize the results of the simulation study:

1. The data was explicitly generated with a multilevel structure of the effects of the unit-covariate Z: The effects of Z_b and Z_w and their product variable on the true-outcome variable τ_0 and the true-effect variable δ_{10} differed considerably [see also Section 2.3 and Equations (4.32) and (4.33)]. This put the naive implementation of the adjustment model in `lace` at a disadvantage, since it only included regression coefficients for the unit-covariate Z. Predictably, the *ACE*-estimator showed a strong and perplexing pattern of bias. If the effects and interactions of the within- and between-cluster components of the unit-covariate Z had been the same, the naive model implementation in `lace` would have yielded an unbiased *ACE*-estimator, but not necessarily the correct standard errors.

2. The data generation procedure also included residual variance components at the cluster-level $Var(r_{0;C})$ and $Var(r_{10;C})$ [see Equations (4.28) and (4.29)] that captured residual effects of the cluster variable C on the true-outcome variable τ_0 and the true-effect variable δ_{10} over and above the influences of the between-component Z_b and the within-component Z_w. This put the implementation of the full adjustment model in `lace` at a disadvantage, since it did not include parameters that would have captured these variance components and was thus likely to yield negatively biased standard errors (Moerbeek et al., 2000, 2001; Raudenbush & Liu, 2000; Snijders & Bosker, 1999). If the two variances $Var(r_{0;C})$ and $Var(r_{10;C})$ had been zero after controlling for all covariates, i.e., if being in a cluster c had not made the true-outcomes and true-effects more similar after controlling for unit- and cluster-covariates, the singlelevel implementations of the adjustment model in `lace` would have also yielded correct standard errors for the average causal effect (Snijders & Bosker, 1999).

3. At data generation, the true values of Z_b, the expected values $E(Z|C{=}c)$, were used as cluster-covariates, while only the fallible cluster means of Z were available as predictors in the data sets. This should have put all methods at a disadvantage since none of them explicitly corrected the bias in the estimated regression parameters due to the unreliability of the cluster means. Surprisingly, the model implementations in `nlme` and `Mplus` and to a lesser extent in `lace` exhibited only

a small bias in the *ACE*-estimators. However, this finding does not necessarily contradict the theoretical derivations by Lüdtke et al. (2008) and Snijders and Bosker (1999), who showed that the bias introduced by the unreliability of the fallible cluster means of the unit-covariate Z as measures of the values of the regression Z_b is stronger with small clusters and small $ICC(Z)$-values [see also Equation (2.14)]. In order to represent realistic multilevel designs with treatment assignment at the unit-level, we only considered relatively large cluster sizes in the simulation design. This might have been partially responsible for the robustness of the models with fallible measures of the cluster-covariate Z_b. Actually, even in the condition with the smallest cluster sizes and the smallest *ICC* of the unit-covariate Z, the reliability of the cluster-means was 0.725 and did not lead to a considerable bias of the *ACE*-estimator. While this robustness under realistic conditions is reassuring, it does not preclude that a bias could emerge with smaller average cluster sizes and further simulations to study this potential bias are called for.

4. Finally, by repeatedly sampling from the single-unit trial, the realized values of the unit-covariate Z varied from sample to sample, making Z_b and Z_w stochastic predictors. This put the conventional linear mixed model implementation in `nlme` at a disadvantage, since it explicitly assumes fixed predictors that are constant over replications of the simulation (Pinheiro & Bates, 2000). Consequently, and in line with the derivations by Chen (2006) and Kröhne (2009), this model implementation exhibited a negatively biased standard error in conditions with strong interactions between Z_b and the treatment variable X. An extended discussion of modeling covariates as fixed or stochastic predictors in the specification of the generalized ANCOVA in applications is foregone until the general discussion in Chapter 6.

Conventional Multilevel ANCOVA

The conventional multilevel ANCOVA (Moerbeek et al., 2001; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) that does not include interactions between the treatment variable X and the covariates at the unit- and at the cluster-level and is usually discussed in the context of randomized designs as a means to increase the precision of estimation was not tested in the simulation study and compared to the generalized

ANCOVA model. Additionally including this model and its implementation in the different statistical frameworks would have made an already complex simulation design even more complicated and would likely not have yielded further insights above the known analytical and simulation results: It is well-known from singlelevel ANCOVA that a model erroneously specified without interactions does not identify the average causal effect if interactions are present and the conditional effect function is not a constant (Kröhne, 2009; Flory, 2008; Rogosa, 1980). A conventional ANCOVA would have given an adequate estimator of the average causal, if no interactions between the treatment variable X and the covariates had been present. While there were conditions without an interaction of the treatment variable X and the between-component Z_b, the interaction of the treatment variable X and the within-component Z_w was constant across the conditions of the simulation design and always different from zero. Hence, unbiased performance of the conventional ANCOVA would not have been expected. Although the robustness of the conventional ANCOVA for non-randomized multilevel designs with treatment assignment at the unit-level (and its implementations in statistical models) remains a topic for further research, the use of the generalized ANCOVA, that always correctly identifies the average causal effect and includes the conventional ANCOVA without interactions as a special case, is recommended.

Variance Heterogeneity

Although the residuals of the conditional effect function had variances larger than zero [$Var(\upsilon_{10;U}) > 0$ and $Var(r_{10;C}) > 0$], these slightly heterogeneous variances on the unit- and the cluster-level did not result in a bias of the *ACE*-estimators or their standard errors. The bias in the standard error of the `nlme` implementation was due to treating predictors as fixed not to the omission of an additional variance component — a small exploratory simulation study that allowed for heterogeneous unit-level variances in `nlme` yielded almost identically biased standard errors. In a similar vein, the standard errors of the singlegroup multilevel manifest variable model in `Mplus` yielded unbiased standard errors, although the variance heterogeneity was not explicitly modeled. These findings are in line with Korendijk, Maas, Moerbeek, and Van der Heijden (2008) who also found no effects of misspecified variance heterogeneity at the cluster-level on the fixed effects in a conventional hierarchical linear model. Unfortunately, this finding could not be corroborated in the present study due to the fact that no multi-

group multilevel model could be specified in `Mplus`. Nevertheless, results from simulation studies of the singlelevel generalized ANCOVA (Kröhne, 2009) indicate that larger variance heterogeneity between the treatment groups and unequal treatment group sizes could make parameter estimates and standard errors inconsistent, if they are obtained from implementations that do not model this heterogeneity explicitly. The amount and consequences of a larger variance heterogeneity at both the unit- and the cluster-level remain open questions worthy of further studies.

Further Research Needs

Although the simulation design resulted in a large simulation study with realistic conditions, there were some notable omissions that need to be addressed in further studies: First, only a single unit-covariate Z decomposable into its between-component Z_b and its within-component Z_w was included in the simulation design. Although the simulation study demonstrated the considerable complexities of estimating the *ACE* with the generalized ANCOVA even in this relatively simple constellation, it does not speak to the additional complexities and sample size requirements involved in specifying and estimating a model with more than one covariate at the unit- or at the cluster-level. In this case, the correct specification of interactions between the covariates becomes a critical and complicated issue. This is problematic insofar as it is unlikely that controlling a univariate covariate at the unit-level and the cluster-level will ever suffice to achieve unbiasedness of $E(Y \mid Z, V, Z_b)$ in quasi-experimental multilevel designs with treatment assignment at the unit-level. However, in designs with conditional randomization of units to treatment conditions, randomized assignment conditional on a single unit- and cluster-covariate is possible.

A second major shortcoming of the present simulation study is the fact that the implementations of the adjustment models were only compared under the null hypothesis of no average causal effect. While the correct estimation of parameters and their standard errors under the null hypothesis are important for every statistical model to guarantee appropriate tests of statistical significance, they are not sufficient for final conclusions about the applicability and usefulness of a statistical procedure. Especially in the planning of evaluation studies, the power of a design and the statistical analysis for detecting a treatment effect of a certain magnitude are of major interest (Moerbeek et al., 2000; Raudenbush & Liu, 2000). The present study speaks to these issues only insofar, as

those methods that yielded biased *ACE*-estimators and standard errors even under the null hypothesis are clearly not recommendable.

4.4.3 Example Analysis

The example analysis of the NELS:1988 (Curtin et al., 2002) data set was intended to illustrate the performance of various implementations of the generalized ANCOVA with an empirical example that was structurally similar to the simulation study. Although the results of these analyses — a moderately average effect of participating in academic afternoon clubs on science competence — cannot be interpreted causally without the additional assumption of conditional unbiasedness, they were illustrative of the implementations of the adjustment models and the complexities involved in interpreting effect estimates from different statistical models in practice. To obtain a truly unbiased estimator of the average causal effect of the treatment, it is likely that more covariates would have to be considered, e.g., the gender of the students on the unit-level, or the socio-economic status at the unit- and school-level.

The different statistical implementations behaved more or less as expected from the simulation: The numerical estimates from the implementations of the full adjustment model were very close to each other. The standard errors also behaved as expected: `Mplus` yielded the largest standard error, followed by `nlme` and `lace`. The differences between the estimated standard errors were small, due to the small effect size of the interaction between the treatment variable and the between-component of the pre-test and the small residual intraclass correlation coefficient *rICC* of the outcome variable. Consequently, the statistical inferences that could be drawn from the three implementations of the full adjustment model were similar. The analyses also showed that adjusting for the pre-test and decomposing the pre-test into its between- and within-component was called for: The unadjusted average treatment effect and the average effect obtained from the naive model implementation in `lace` differed markedly from the average effect estimates obtained from the full adjustment models.

In judging the adequacy of the average effect estimates from the models against the background of the simulations study, two caveats need to be taken into account: (1) The simulation only assessed the appropriateness of the statistical models under the null hypothesis of no average effect. In the empirical example, however, the average treatment effect after controlling for the pre-test was significantly positive. There is no

information to be gained from the simulation study that guarantees that the properties of the statistical methods also hold under these conditions. (2) The average reliability of the cluster means was considerably smaller in the empirical example than in any condition of the simulation study. Clearly, the observed cluster-means of the pre-test were not equal to the true school means of the pre-test, since only a sample of students from each school was included in the study. However, it is likely that the true school means of the pre-test and not the observed school means were the covariate that in fact influenced the treatment assignment probabilities and the outcome variable. Methods that correct for the unreliability of the cluster means could potentially lead to different *ACE*-estimates. The development of statistical models that correct this unreliability for designs with treatment assignment at the unit-level is a fruitful area of further research.

4.4.4 Designs with Unbiasedness of $E(Y \mid X, Z, C)$

Statistical implementations of the generalized ANCOVA for designs that do not lead to unbiasedness of $E(Y \mid X, Z, V, Z_b)$ but only fulfill the weaker condition of unbiasedness of $E(Y \mid X, Z, C)$ were not considered in detail. As discussed earlier, estimation of the average causal effect in applications requires the estimation of the expected value of the product of a function of the cluster variable C and the unit-covariate Z that depends on the covariance of the cluster variable C and the unit-covariate Z [see Equation (4.17)]. We will briefly outline the options for statistical models to implement such adjustment models that could be studied in further research.

One obvious option to estimate the *ACE* in designs in with unbiasedness of the regression $E(Y \mid X, Z, C)$ is to parametrize the cluster variable C with indicator variables and model the interaction with the unit-covariate Z with the products of the indicators and Z. Such an approach is also known as a fixed-effects model in conventional multilevel terminology (Gelman & Hill, 2007; Snijders & Bosker, 1999). Unfortunately, the number of regression weights that have to be estimated is considerably large, if many clusters are considered and the computation and specification of the expected value of the average effect functions becomes cumbersome.

The second option is to estimate the generalized ANCOVA with a hierarchical linear regression model decomposing the cluster-specific functions $f_2(C)$ and $f_3(C)$ into their expected value and a corresponding residual. This approach will provide estimates of the expected values of the cluster-specific functions directly, but the statistical model

implementation rests on the assumption that the cluster-specific deviations of $f_2(C)$ and $f_3(C)$ from their respective expected values are uncorrelated with the other predictors in the model — otherwise the resulting parameter estimates will be inconsistent (Kim & Frees, 2006, 2007; Raudenbush & Bryk, 2002). This condition is only fulfilled, if $Cov[f_3(C), Z]$ is equal to zero. As discussed earlier, this condition is guaranteed to hold in designs with randomized assignment of units to clusters, but can be violated if the *ICC* of the unit-covariate Z is larger than zero.

One alternative is to further specify the conditional effect function by including predictors on the cluster-level, especially the between-component Z_b and specifying cross-level interaction between these predictors and the unit-covariate Z. The resulting statistical models would be close to the models studied in the simulation study, but would include further random effects and residual components to account for the interaction of the cluster variable C and other predictors in the models. There are two complications to this approach: (1) The use of additional predictors requires the specification the functional form of $f_3(C)$ and thus introduces another potential source of bias. (2) Unbiasedness of $E(Y \mid X, Z, C)$ does not imply unbiasedness of $E(Y \mid X, Z, V, Z_b)$. However, it is exactly this regression that would be modeled implicitly.

Finally, there are emerging methods based on instrumental variables approaches to model estimation, that give consistent estimates for regression parameters even if there is a correlation between the predictors and residual terms (Kim & Frees, 2006, 2007). The options for implementing the generalized ANCOVA with these approaches and their use to estimate average causal effects deserve further research.

4.4.5 Recommendations and Conclusion

Considering the results of the simulation as a whole, the outlook for implementing the generalized ANCOVA for designs with treatment assignment at the unit-level in statistical models and for applications is not too dismal under the conditions considered in the simulation study: The multilevel model implementation in `Mplus` that treated predictors as stochastic yielded unbiased *ACE*-estimates and (mostly) correct standard errors, when it converged. It can thus be recommended for applications of the adjustment models in sample sizes similar to the ones realized in the simulation study. The conventional hierarchical linear regression model in `nlme` potentially underestimates the standard error of the *ACE*-estimator, if predictors are stochastic and there are in-

teractions of the between-component Z_b and the treatment variable X — uncritically using this model in applications can thus not be recommended. Alternative implementations of the generalized ANCOVA that take the potential bias due to unreliability of the cluster means as predictors into account (Croon & van Veldhoven, 2007; Grilli & Rampichini, 2008; Lüdtke et al., 2008) still need to be developed and may be considered in further studies.

While we have focused on the statistical implementations and the simulation study in the preceding section, the general appropriateness of the generalized ANCOVA for causal inferences and other overarching problems will be given further attention in the general discussion in Chapter 6.

5 Average Causal Effects for Treatment Assignment at the Cluster-Level

This chapter discusses the analysis of causal effects in multilevel designs with treatment assignment at the cluster-level. The statistical analysis of cluster randomized trials is well understood (Donner & Klar, 2000; Murray, 1998) and average causal effects are identified with prima-facie effects in cluster randomized designs. Hence, we will focus on designs in which unbiasedness of the treatment regression $E(Y \mid X)$ does not hold. Specifically, we will develop an adjustment procedure that extends the generalized analysis of covariance (ANCOVA) introduced by Steyer et al. (2009) to conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the cluster-level. The identification of average causal effects in these designs is not as well understood as causal inference for unconditionally cluster randomized designs. Conventional accounts of the multilevel ANCOVA for designs with treatment assignment at the cluster-level (Bloom et al., 1999, 2007; Donner & Klar, 2000; Gitelman, 2005; Moerbeek et al., 2001; Murray, 1998; Oakes, 2004; Raudenbush, 1997; Raudenbush et al., 2007; Seltzer, 2004; Pituch, 2001; Plewis & Hurry, 1998; VanderWeele, 2008) do not identify the average causal effect in the presence of interactions between the treatment variable and the covariates.

The chapter is structured as follows: We first identify average causal effects for general effect functions. Next, we turn to linear effect functions and develop a generalized ANCOVA for non-randomized designs with treatment assignment at the cluster-level. More specifically, we develop two adjustment models — one that includes only covariates at the cluster-level, the second one also including covariates at the unit-level — and show how the average causal effect can be identified as a non-linear function of the

parameters of a multiple linear regression in both cases. Then, we report the results of a simulation study that compared the finite sample performance of several statistical implementations of the adjustment models under the null hypothesis of no average causal effect in a design with a unit-covariate Z and the corresponding between-component Z_b. Finally, we illustrate the statistical models with an empirical example from the Early Childhood Longitudinal Study - Kindergarten Class of 1998-99 (ECLS-K, National Center for Education Statistics, 2001). Throughout this chapter, we always refer to the multilevel causality space $\langle(\Omega, \mathfrak{A}, P), (\mathfrak{C}_t)_{t\in T}, X, Y, \mathfrak{C}_X\rangle$ introduced in Chapter 2. Our discussion pertains to designs with assignment of units to clusters as well as designs that use pre-existing clusters.

5.1 Adjustment Models

In this section, we develop the generalized ANCOVA (Steyer et al., 2009) for designs in which whole clusters are assigned to different treatment conditions with differing probabilities, more specifically for conditionally randomized and quasi-experimental designs. In these designs, the true mean differences of the outcome variable between treatment conditions, in general, do no longer identify average causal effects. In the remainder of the section, we will consider designs in which unbiasedness of the cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$ holds [for the definition see Equation (2.56)]. In experimental designs, unbiasedness of $E(Y \mid X, V, Z_b)$ is implied when all clusters with values v of the cluster-covariate V and z_b of the between-component Z_b have the same probabilities of being assigned to treatment and control conditions [see Equation (3.9)]. In quasi-experimental designs, unbiasedness of $E(Y \mid X, V, Z_b)$ does not hold automatically, but requires the inclusion of all cluster-covariates V and between-components Z_b that influence both the outcome variable Y and the treatment assignment probabilities.

In multilevel designs with treatment assignment at the cluster-level, the unit-covariate Z cannot further confound the regression $E(Y \mid X, V, Z_b)$. Since whole clusters are assigned to treatment conditions — the treatment variable X is conceptually a variable at the cluster-level — treatment assignment probabilities can only depend on cluster-covariates V and the between-component Z_b, but not on the unit-covariate Z. Stated more formally, this implies:

$$P(X{=}j \mid Z, V, Z_b) = P(X{=}j \mid V, Z_b) \quad \text{a.s. for all values } j \text{ of } X. \tag{5.1}$$

In other words, the unit-covariate Z can only influence treatment assignment probabilities by its between-component Z_b (or, more generally, by any other function of its cluster-specific distribution that can be represented as a cluster-covariate). Once the between-component Z_b is taken into account, the treatment assignment probabilities are independent of the unit-covariate Z and of the within-component Z_w. Equation (5.1) is structurally similar to the first part of the second definition of conditional unconfoundedness (Steyer et al., 2009) and implies that the unit-covariate Z does not confound $E(Y \mid X, V, Z_b)$ (for the proof, see, Steyer et al.). More specifically, if the regression $E(Y \mid X, V, Z_b)$ is unbiased, Equation (5.1) implies that average stability of the regression $E(Y|X, V, Z_b)$ holds with respect to unit-covariate Z (the proof is given in Appendix A.1):

$$E^{\circ}_{X=j}(Y \mid V, Z_b) = E\big[E^{\circ}_{X=j}(Y \mid Z, V, Z_b) \mid V, Z_b\big]. \tag{5.2}$$

The values of the extensions of the conditional regressions $E_{X=j}(Y \mid V, Z_b)$ can always be obtained by averaging over the extensions of the regressions $E_{X=j}(Y|Z, V, Z_b)$ conditional on the cluster-covariate V and the between-component Z_b. As we will show in the next section, this implies that adjustment models, such as the generalized ANCOVA, can be specified using either the regression $E(Y \mid X, V, Z_b)$ or the regression $E(Y \mid X, Z, V, Z_b)$ — the resulting ACE_{jk}-identifiers will be identical (see also, VanderWeele, 2008).

However, it might be advantageous or even necessary to include the unit-covariate Z in statistical implementations of the generalized ANCOVA for two reasons: (1) If the unit-covariate Z influences the outcome variable Y over and above the between-component Z_b, including Z in a statistical model reduces the residual error variance and thus increases efficiency of the ACE-estimator and power to detect treatment effects (Murray, 1998; Raudenbush, 1997; VanderWeele, 2008). (2) The between-component Z_b is, in fact, a latent variable whose values are only approximated by the empirical cluster means in applications. Using the cluster means of the unit-covariate Z without further corrections will lead to biased ACE-estimator in applications for small clusters [see Equation (2.14)]. Some statistical models that correct this unreliability, e.g., the multigroup multilevel latent variable model in `Mplus` 5.0 (B. O. Muthén, 1998-2004), require the explicit inclusion of the unit-covariate Z (or the within-component Z_w) as predictor.

In the next sections, we show how the assumption of unbiasedness of $E(Y \mid X, V, Z_b)$ allows to identify average causal effects with empirically estimable quantities using ei-

ther the regression $E(Y|X, V, Z_b)$ or the regression $E(Y|X, Z, V, Z_b)$. We will first develop the adjustment model for general effect functions and then turn to effect functions that are linear in the covariates and their products.

5.1.1 General Effect Functions

In this section, we will develop the adjustment models for general effect functions under the assumption of unbiasedness of $E(Y \mid X, V, Z_b)$. We will first develop the simple adjustment model, using the regression $E(Y|X, V, Z_b)$ and then introduce the full adjustment model that additionally includes the unit-covariate Z and refers to the regression $E(Y \mid X, Z, V, Z_b)$. We will show that both adjustment models yield equivalent ACE-identifiers.

Simple Adjustment Model

We begin with discussing identification of the average causal effect ACE_{jk} in non-randomized multilevel design with treatment assignment at the cluster-level for designs in which the cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$ is unbiased. Unbiasedness of $E(Y \mid X, V, Z_b)$ has been defined in Equation (2.56) — for convenience this definition is repeated here, explicitly including the between-component Z_b among the random variables for which unbiasedness is required:

$$E^{\circ}_{X=j}(Y \mid V, Z_b) = E(\tau_j \mid V, Z_b) \quad \text{a.s. for all values } j \text{ of } X.$$

If the $(J + 1)$-valued treatment variable X is represented with J dummy variables $I_{X=j}$ with values 0 and 1, one for each treatment condition using the control condition as reference, $E(Y \mid X, V, Z_b)$ can always be written as:

$$E(Y \mid X, V, Z_b) = g_0(V, Z_b) + g_1(V, Z_b) \cdot I_{X=1} + \ldots + g_J(V, Z_b) \cdot I_{X=J}. \tag{5.3}$$

Without further assumptions, the function $g_0(V, Z_b)$ is the regression $E_{X=0}(Y \mid V, Z_b)$ of the outcome variable Y on the cluster-covariate V and the between-component Z_b in the control group. The functions $g_j(V, Z_b)$ are the conditional prima-facie effect functions $PFE_{j0\,;\,V,Z_b}$ whose values are the conditional prima-facie effects of treatment condition j compared to the control condition for all combinations of values of the cluster-covariate

V and the between-component Z_b as defined in Equation (2.49)

$$g_j(V, Z_b) = E^{\circ}_{X=j}(Y \mid V, Z_b) - E^{\circ}_{X=0}(Y \mid V, Z_b) = PFE_{j0\,;\,V,Z_b}. \tag{5.4}$$

Under unbiasedness of $E(Y \mid X, V, Z_b)$, the values of the $g_j(V, Z_b)$-functions are not only prima-facie effects, but also conditional causal effects; the $g_j(V, Z_b)$-functions are equal to the conditional causal effect functions $CCE_{j0\,;\,V,Z_b}$ as defined in Equation (2.38):

$$g_j(V, Z_b) = E(\tau_j \mid V, Z_b) - E(\tau_0 \mid V, Z_b) = CCE_{j0\,;\,V,Z_b}. \tag{5.5}$$

As shown in Equation (2.39), the expected value of the conditional causal effect function $E(CCE_{jk\,;\,V,Z_b})$ is equal to the average causal effect ACE_{jk}. Since the conditional causal effect function $CCE_{j0\,;\,V,Z_b}$ is equal to $g_j(V, Z_b)$ under unbiasedness of the cluster-covariate-treatment regression $E(Y \mid V, Z_b)$, the expected value $E[g_j(V, Z_b)]$ is equal to the average causal effect ACE_{j0}.

$$ACE_{j0} = (CCE_{jk\,;\,V,Z_b}) = E[g_j(V, Z_b)]. \tag{5.6}$$

This equality can be used to identify average causal effects in non-randomized multilevel designs with treatment assignment at the cluster-level. The practical identification from samples requires that the functional forms of the intercept function $g_0(V, Z_b)$ and the effect functions $g_j(V, Z_b)$ are estimated (Steyer et al., 2009) — either by specifying their parametric form or by modeling them non-parametrically.

Full Adjustment Model

We will now develop the full adjustment model that includes the unit-covariate Z and uses the regression $E(Y \mid X, Z, V, Z_b)$ with general effect functions. Since $E(Y \mid X, V, Z_b)$ is unconfounded with respect to the unit-covariate Z, the resulting ACE-estimators are equivalent, but modeling the unit-covariate Z explicitly is useful in some statistical models. Hence, in order to use the regression $E(Y \mid X, Z, V, Z_b)$ to identify the average causal effect, we only have to assume that the regression $E(Y \mid X, V, Z_b)$ is unbiased (see Appendix A.1).

If the $(J + 1)$-valued treatment variable X is represented with J dummy variables $I_{X=j}$ with values 0 and 1, one for each treatment condition using the control condition as

reference, $E(Y \mid X, Z, V, Z_b)$ can always be written as:

$$E(Y \mid X, Z, V, Z_b) = g_0(Z, V, Z_b) + g_1(Z, V, Z_b) \cdot I_{X=1} + \ldots + g_J(Z, V, Z_b) \cdot I_{X=J}. \quad (5.7)$$

The function $g_0(Z, V, Z_b)$ is the regression $E_{X=0}(Y \mid Z, V, Z_b)$ of the outcome variable Y on the unit-covariate Z, the cluster-covariate V and the between-component Z_b in the control group without further assumptions. In the same vein, the functions $g_j(Z, V, Z_b)$ are the differences between the extensions of the regressions of the outcome variable Y on the unit-covariate Z, the cluster-covariate V and the between-component Z_b in treatment group j and the control group respectively

$$g_j(Z, V, Z_b) = E^{\circ}_{X=j}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b) = PFE_{j0\,;\,Z,\,V,\,Z_b}. \quad (5.8)$$

The difference between these two extensions is the conditional prima-facie effect function $PFE_{j0\,;\,Z,\,V,\,Z_b}$ whose values are the conditional prima-facie effects of treatment condition j compared to the control condition for all combinations of values of the unit-covariate Z, the cluster-covariate V and the between-component Z_b as introduced in Equation (2.52)

Average stability of the regression $E(Y \mid X, V, Z_b)$ with respect to the unit-covariate Z as introduced in Equation (5.2) — that holds in all multilevel designs with treatment assignment at the cluster-level since the unit-covariate Z cannot influence the treatment assignment probabilities over and above the cluster-covariate V and the between-component Z_b — implies that

$$\begin{aligned} E[g_j(Z, V, Z_b) \mid V, Z_b] &= E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b) \mid V, Z_b] \\ &= E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b) \mid V, Z_b] - E[E^{\circ}_{X=0}(Y \mid Z, V, Z_b) \mid V, Z_b] \\ &= E^{\circ}_{X=j}(Y \mid V, Z_b) - E^{\circ}_{X=0}(Y \mid V, Z_b) \\ &= g_j(V, Z_b). \end{aligned} \quad (5.9)$$

Under unbiasedness of $E(Y \mid X, V, Z_b)$, the values of $g_j(V, Z_b)$ are not only prima-facie effects, but also conditional causal effects and the $g_j(V, Z_b)$-functions are equal to the conditional causal effect functions $CCE_{j0\,;\,V,\,Z_b}$ as defined in Equation (2.40). The expected value $E[g_j(V, Z_b)]$ then identifies the ACE_{j0}.

In multilevel designs with treatment assignment at the cluster-level, the following

equality holds (for the proof see Appendix A.1):

$$E[g_j(Z, V, Z_b)] = E[E[g_j(Z, V, Z_b) \mid V, Z_b]] = E[g_j(V, Z_b)] \tag{5.10}$$

The expected value $E[g_j(Z, V, Z_b)]$ is always equal to $E[E[g_j(Z, V, Z_b) \mid V, Z_b]]$. Hence, the expected value of the conditional effect function $E[g_j(V, Z_b)]$ from the simple adjustment model is always equal to the expected value of the conditional effect function $E[g_j(Z, V, Z_b)]$ from the full adjustment model (for the proof see Appendix A.1). This implies that the average causal effect is also identified with $E[g_j(Z, V, Z_b)]$ under unbiasedness of $E(Y \mid X, V, Z_b)$. This equality can be used to identify average causal effects in non-randomized multilevel designs with treatment assignment at the cluster-level. Again, in order to do this, the functional forms of the intercept function $g_0(Z, V, Z_b)$ and the effect functions $g_j(Z, V, Z_b)$ have to be estimated (Steyer et al., 2009) — either by specifying their parametric form or by modeling them non-parametrically.

5.1.2 Linear Effect Functions

The cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$ and the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, Z, V, Z_b)$ were introduced in Equations 5.3 and 5.7 without further assumptions about the respective intercept functions g_0 and effect functions g_j. If the adjustment methods are used in applications, the functional forms of g_0 and g_j must be explicitly specified or alternatively modeled non-parametrically (Steyer et al., 2009). As a consequence, the validity of causal inferences in applications depends on the correct specification of the regressions of the outcome variable Y on the set of considered covariates in each treatment group $E_{X=j}(Y \mid V, Z_b)$ or $E_{X=j}(Y \mid Z, V, Z_b)$ respectively. Any misspecification of these regressions can result in a severe bias of the estimated average causal effect (see also D. J. Bauer & Cai, 2008; Kang & Schafer, 2006).

In the remainder of this section, we develop the generalized ANCOVA (Kröhne, 2009; Steyer et al., 2009) for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the cluster-level using linear intercept and effect functions (see also Kröhne, 2009, for a similar formulation with regard to the parametrization of the regressions of the outcome variable Y on the covariate Z for each treatment group j). We start with designs that lead to an unbiased cluster-

covariate-treatment regression $E(Y \mid X, V, Z_b)$ and develop the *simple adjustment model* based on this regression. Then, we will develop the *full adjustment model* based on $E(Y \mid X, Z, V, Z_b)$. If a linear parametrization for intercept and effect functions is chosen, the validity of causal inferences is contingent not only on the assumption of unbiasedness of the respective regressions, but also on the tenability of the linearity assumptions for the functions g_0 and g_j.

The generalized ANCOVA, introduced in the next section, extends and consolidates the different versions of multilevel ANCOVA models that have been proposed for multilevel designs with treatment assignment at the cluster-level (Bloom et al., 1999, 2007; Donner & Klar, 2000; Gitelman, 2005; Moerbeek et al., 2001; Murray, 1998; Oakes, 2004; Raudenbush, 1997; Raudenbush et al., 2007; Seltzer, 2004; Pituch, 2001; Plewis & Hurry, 1998; VanderWeele, 2008) in the following ways:

1. It explicitly acknowledges the decomposition of the unit-covariate Z into the within-component Z_w and the between-component Z_b as did Bloom et al. (1999, 2007), Moerbeek et al. (2001), Oakes (2004) and Raudenbush (1997).

2. It includes interactions between the covariates Z, V, Z_b and the treatment variable X (Pituch, 2001; Plewis & Hurry, 1998; Seltzer, 2004), but yet identifies an average treatment effect (see also Flory, 2008; Kröhne, 2009; Nagengast, 2006; Rogosa, 1980) and not only conditional treatment effects and details and corrects the implicit assumptions made in similar endeavours by Gitelman (2005) and VanderWeele (2008).

3. It is embedded in an explicit theory of causality and uses covariates explicitly to identify the average causal effect ACE_{jk} in conditionally-randomized and quasi-experimental designs (see also Gitelman, 2005; VanderWeele, 2008). Conventionally, covariates have often been discussed only as means to improve precision in randomized designs (Bloom et al., 2007; Donner & Klar, 2000; Moerbeek et al., 2001; Murray, 1998; Raudenbush, 1997; Raudenbush et al., 2007).

Simple Adjustment Model

Our discussion of the adjustment model for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the unit-level that lead to unbiasedness of $E(Y|X, V, Z_b)$ will be confined to a binary treatment variable X, representing

a treatment and a control condition. We will only consider one cluster-covariate V and one cluster-component Z_b as covariates. Generalizations to more than two treatment conditions and more covariates are straightforward, but require further assumptions about the interactions between the covariates.

If the regression $E(Y \mid X, V, Z_b)$ is linear in the treatment variable X, the cluster-covariate V, the between-component Z_b and their products, it can be written as:

$$\begin{aligned} E(Y \mid X, V, Z_b) = \gamma_{00} + \gamma_{01} V + \gamma_{02} Z_b + \gamma_{03} V \cdot Z_b \\ + \left[\gamma_{10} + \gamma_{11} V + \gamma_{12} Z_b + \gamma_{13} V \cdot Z_b\right] \cdot X. \end{aligned} \tag{5.11}$$

The outcome variable Y can be decomposed into the regression $E(Y \mid X, V, Z_b)$ and its residual $\varepsilon \equiv Y - E(Y \mid X, V, Z_b)$

$$Y = E(Y \mid X, V, Z_b) + \varepsilon. \tag{5.12}$$

The residual ε has all properties of a residual of a multiple linear regression, most notably its expected value and its regression on the regressors is zero. The last property only holds, if the regression of the outcome variable Y on the regressors X, V, Z_b and their products is in fact linear. Otherwise, only the properties of the residual of a linear ordinary least-squares regression $Q(Y \mid X, V, Z_b)$ hold (e.g., Kutner et al., 2005; Rechner & Schaalje, 2007, Chapter 8). No further assumptions about the distribution of ε are made at this point. However, these assumptions, e.g., about the homogeneity of variances between treatment groups or about dependencies within clusters, distinguish the statistical models used for the implementation of the adjustment model in different statistical frameworks in the simulation study in Section 5.2.

Next, we derive the parameters of the conditional effect function $g_1(V, Z_b)$ from the parameters of Equations (5.11). In Equation (5.4), we noted that $g_1(V, Z_b)$ is equal to the difference of the extensions of the conditional regressions $E_{X=1}(Y \mid V, Z_b)$ and $E_{X=0}(Y \mid V, Z_b)$. Taking the parameters of in Equation (5.11), these two regressions have the following form:

$$E_{X=0}(Y \mid V, Z_b) = \gamma_{00} + \gamma_{01} V + \gamma_{02} Z_b + \gamma_{03} V \cdot Z_b \tag{5.13}$$

$$\begin{aligned} E_{X=1}(Y \mid V, Z_b) = \gamma_{00} + \gamma_{10} + (\gamma_{01} + \gamma_{11}) V + (\gamma_{02} + \gamma_{12}) Z_b \\ + (\gamma_{03} + \gamma_{13}) V \cdot Z_b. \end{aligned} \tag{5.14}$$

Equation (5.13) describes the extension of the regression $E_{X=0}(Y \mid V, Z_b)$ of the outcome variable Y on the covariates in the control condition and is thus equal to the intercept function $g_0(V, Z_b)$. The conditional effect function $g_1(V, Z_b)$ is obtained by subtracting the extension of Equation (5.13) from the extension of Equation (5.14):

$$\begin{aligned} g_1(V, Z_b) &= E^{\circ}_{X=1}(Y \mid V, Z_b) - E^{\circ}_{X=0}(Y \mid V, Z_b) \\ &= \gamma_{10} + \gamma_{11} V + \gamma_{12} Z_b + \gamma_{13} V \cdot Z_b. \end{aligned} \tag{5.15}$$

To obtain the average causal effect ACE_{10}, the expected value of Equation (5.15) has to be taken [see also Equation (5.6)] and the algebraic rules for expected values and regressions (cf., H. Bauer, 1981) have to be applied to the results:

$$\begin{aligned} E[g_1(Z_b, V)] &= E[\gamma_{10} + \gamma_{11} V + \gamma_{12} Z_b + \gamma_{13} V \cdot Z_b] \\ &= \gamma_{10} + \gamma_{11} E(V) + \gamma_{12} E(Z) + \gamma_{13} E(V \cdot Z_b). \end{aligned} \tag{5.16}$$

The average causal effect ACE_{10} in experimental multilevel designs with conditional randomization or quasi-experimental designs with treatment assignment at the cluster-level, assuming linear intercept- and effect-functions and unbiasedness of the cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$, is given by the following non-linear function of model parameters and expected values of the covariates and their product variable

$$ACE_{10} = \gamma_{10} + \gamma_{11} E(V) + \gamma_{12} E(Z) + \gamma_{13} E(V \cdot Z_b). \tag{5.17}$$

As in singlelevel models (Kröhne, 2009; Steyer et al., 2009), the ACE_{10} is identified as a non-linear function of regression parameters and expected values of the covariates and their product. If there are no interactions between the treatment variable X and the covariates, i.e., if the corresponding regression weights γ_{11}, γ_{12} and γ_{13} of the interactions between the treatment variable and the corresponding covariates are *all* equal to zero and, hence, the conditional effect function $g_1(V, Z_b)$ is a constant, the average causal effect ACE_{10} is identified by the regression parameter γ_{10} of the treatment indicator variable. This is in line with presentations of the conventional multilevel ANCOVA in general (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and for designs with treatment assignment at the cluster-level specifically (Bloom et al., 1999, 2007; Moerbeek et al., 2001; Oakes, 2004; Raudenbush, 1997) where no interactions between the

treatment variable and the covariates are included. It is well-known from singlelevel ANCOVA that the estimate from a model without interactions does not identify the average causal effect ACE_{10} in the presence of non-zero interaction effects, but only the conditional treatment effect at the point of highest precision [see Rogosa, 1980, Eqs. (18) and (19)]. Hence, the conventional ANCOVA without interactions cannot be recommended for identification of the average causal effect without qualifications.

If the expected values of the covariates and their product variable — $E(Z)$, $E(V)$ and $E(Z_b \cdot V)$ — are equal to zero, the average causal effect ACE_{10} is also simply identified by the regression parameter γ_{10} of the treatment indicator variable. Centering covariates by subtracting their empirical means from the observed values in applications before calculating the product terms with the treatment variable, makes their empirical mean equal to zero (Aiken & West, 1996; Kreft et al., 1995) — but not necessarily the observed mean of the product that depends on the covariance $Cov\,(Z_b, V)$. While centering covariates around the empirical means identifies the average causal effect with a single parameter — if there are no three-way or higher-order interactions of the covariates and the treatment variable — analytical derivations (Chen, 2006) and simulation studies (Kröhne, 2009; Nagengast, 2006) of statistical implementations of singlelevel adjustment models show that standard errors of the *ACE* are underestimated, if covariates are stochastic predictors in statistical models and not fixed by design. Centering can therefore not be recommended without qualifications for obtaining standard errors of the average causal effect for designs with stochastic covariates and observational studies.

Full Adjustment Model

As we did in the presentation of the simple adjustment model, we will introduce the full adjustment model for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the cluster-level using only a binary treatment variable X, representing a treatment and a control condition, and will only consider one unit-covariate Z, one cluster-covariate V, the between-component Z_b and their products. Generalizations to more than two treatment conditions and more covariates are straightforward, but require further assumptions about the interactions to be considered.

There are two equivalent ways to parametrize the regression $E(Y \mid X, Z, V, Z_b)$: (1) Either with the raw scores of the unit-covariate Z or (2) with the within-component Z_w [as defined in Equation (2.3)]. Both parametrizations give identical information with

respect to the conditional expected values of the outcome variable Y given the treatment variable X and the covariates since the within-component Z_w is defined as the difference between the raw values of the unit-covariate Z and the between-component Z_b (see also Enders & Tofighi, 2007; Kreft et al., 1995), but the meaning of model parameters differ. For the following derivations, the parametrization using the within-component Z_w is used, as it results in a simpler identifier of the average causal effect. At the end of the section, we will briefly discuss the alternative parametrization of $E(Y \mid X, Z, V, Z_b)$ that uses the raw scores of the unit-covariate Z and is relevant for one of the statistical models in the simulation study.

If the regression $E(Y \mid X, Z, V, Z_b)$ is linear in the treatment variable X, the within-component of the unit-covariate Z_w, the cluster-covariate V, the between-component of the unit-covariate Z_b and their products, it can be written as

$$\begin{aligned} E(Y \mid X, Z, V, Z_b) = {} & \gamma_{00} + \gamma_{01} Z_b + \gamma_{02} V + \gamma_{03} Z_b \cdot V \\ & + \left[\gamma_{04} + \gamma_{05} Z_b + \gamma_{06} V + \gamma_{07} Z_b \cdot V\right] \cdot Z_w \\ & + \left[\gamma_{10} + \gamma_{11} Z_b + \gamma_{12} V + \gamma_{13} Z_b \cdot V \right. \\ & \left. + \left[\gamma_{14} + \gamma_{15} Z_b + \gamma_{16} V + \gamma_{17} Z_b \cdot V\right] \cdot Z_w\right] \cdot X. \end{aligned} \tag{5.18}$$

The outcome variable Y can be decomposed into the regression $E(Y \mid X, Z, V, Z_b)$ and its residual $\varepsilon \equiv Y - E(Y \mid X, Z, V, Z_b)$

$$Y = E(Y \mid X, Z, V, Z_b) + \varepsilon. \tag{5.19}$$

Again, the residual ε has all properties of a residual of a multiple linear regression, most notably its expected value and its regression on the regressors is zero. The last property only holds, if the regression of the outcome variable Y on the regressors X, Z_w, Z_b, V and their products is, in fact, linear. Otherwise, only the properties of the residual of a linear ordinary least-squares regression $Q(Y \mid X, Z, V, Z_b)$ hold (e.g., Kutner et al., 2005; Rechner & Schaalje, 2007, Chapter 8). As in the simple adjustment model, no further assumptions about the distribution of ε are made at this point, but will be relevant for the model implementations compared in the simulation study. The assumptions made about ε also distinguish the analysis of designs with treatment assignment at the unit-level — for which a similar regression $E(Y \mid X, Z, V, Z_b)$ was given in Equation (4.18) — from designs with treatment assignment at the cluster-level discussed in this chapter.

Next, we derive the parameters of the conditional effect function $g_1(Z, V, Z_b)$ from the parameters of Equations (5.18). In Equation (5.8), we noted that $g_1(Z, V, Z_b)$ is equal to the difference between the extensions of the conditional regressions $E_{X=1}(Y \mid Z, V, Z_b)$ and $E_{X=0}(Y \mid Z, V, Z_b)$. Using the parameters of Equation (5.18), these two regressions have the following form:

$$\begin{aligned} E_{X=0}(Y \mid Z, V, Z_b) &= \gamma_{00} + \gamma_{01}Z_b + \gamma_{02}V + \gamma_{03}Z_b \cdot V \\ &\quad + [\gamma_{04} + \gamma_{05}Z_b + \gamma_{06}V + \gamma_{07}Z_b \cdot V] \cdot Z_w, \end{aligned} \tag{5.20}$$

$$\begin{aligned} E_{X=1}(Y \mid Z, V, Z_b) &= \gamma_{00} + \gamma_{10} + (\gamma_{01} + \gamma_{11})Z_b \\ &\quad + (\gamma_{02} + \gamma_{12})V + (\gamma_{03} + \gamma_{13})Z_b \cdot V \\ &\quad + [\gamma_{04} + \gamma_{14} + (\gamma_{05} + \gamma_{15})Z_b \\ &\quad + (\gamma_{06} + \gamma_{16})V + (\gamma_{07} + \gamma_{17})Z_b \cdot V] \cdot Z_w. \end{aligned} \tag{5.21}$$

Equation (5.20) describes the extension of the regression $E_{X=0}(Y \mid Z, V, Z_b)$ of the outcome variable Y on the covariates in the control condition and is thus equal to the intercept function $g_0(Z, V, Z_b)$. The conditional effect function $g_1(Z, V, Z_b)$ is obtained by subtracting the extension of Equation (5.20) from the extension of Equation (5.21):

$$\begin{aligned} g_1(Z, V, Z_b) &= E^{\circ}_{X=1}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b) \\ &= \gamma_{10} + \gamma_{11}Z_b + \gamma_{12}V + \gamma_{13}Z_b \cdot V \\ &\quad + [\gamma_{14} + \gamma_{15}Z_b + \gamma_{16}V + \gamma_{17}Z_b \cdot V] \cdot Z_w. \end{aligned} \tag{5.22}$$

To identify the average causal effect ACE_{10}, the expected value of Equation (5.4) has to be taken and simplified by applying the algebraic rules for expected values and regressions (cf. H. Bauer, 1981) to the resulting function :

$$\begin{aligned} E[g_1(Z, V, Z_b)] &= \gamma_{10} + \gamma_{11}E(Z_b) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V) \\ &\quad + \gamma_{14}E(Z_w) + \gamma_{15}E(Z_b \cdot Z_w) \\ &\quad + \gamma_{16}E(V \cdot Z_w) + \gamma_{17}E(Z_b \cdot V \cdot Z_w) \end{aligned} \tag{5.23}$$

$$= \gamma_{10} + \gamma_{11}E(Z) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V). \tag{5.24}$$

Equation (5.23) simplifies considerably to Equation (5.24). These simplifications are possible by taking into account that Z_w is the residual of the regression $E(Z \mid C)$ (see

Section 2.3 for the discussion of this regression and its residual). By virtue of this relation, its expected value is equal to zero [$E(Z_w) = 0$] and it is regressively independent from all functions of its regressor C, making the expected values of its products with Z_b, V and $Z_b \cdot V$ also equal to zero [$E(Z_b \cdot Z_w) = 0$, $E(V \cdot Z_w) = 0$ and $E(Z_b \cdot V \cdot Z_w) = 0$], at least if the cluster-covariate V is a function of the cluster variable C. In case of a cluster-covariate V that is not a function of C, the corresponding parameter γ_{16} of the product of Z_w and V remains in Equation (5.24).

Thus, the average causal effect ACE_{10} in experimental multilevel designs with conditional randomization or quasi-experimental multilevel designs with treatment assignment at the cluster-level assuming linear intercept- and effect-functions, taking separate within- and between-effects of the unit-covariate Z into account and assuming unbiasedness of the unit-covariate-cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$ is given by the following non-linear function of model parameters and expected values of the covariates and their product

$$ACE_{10} = \gamma_{10} + \gamma_{11}E(Z) + \gamma_{12}E(V) + \gamma_{13}E(Z_b \cdot V). \tag{5.25}$$

Equation (5.25) is structurally similar to Equation (5.17) that identified the average causal effect using the regression $E(Y|X, V, Z_b)$. In fact, as discussed in Section 5.1.1, the simple and the full adjustment model lead to identical identifiers of the average causal effect ACE_{j0}, since the regression $E(Y \mid X, V, Z_b)$ is always unconfounded with respect to the unit-covariate Z and the within-component Z_w. In multiple linear regressions, the corresponding partial regression coefficients for the cluster-covariates are identical for the simple and the full adjustment model (Rao, 1973), since the within-component Z_w is regressively independent of the cluster-covariates V and the between-component Z_b, at least if the cluster-covariate V is a function of the cluster-variable C. Hence, the identifier of the ACE_{10} obtained from the full adjustment model in Equation (5.25) is equal to the ACE-identifier from the simple adjustment model given in Equation (5.17). However, the efficiency of the estimator that explicitly uses both unit- and cluster-covariates is higher, if Z_w is a significant predictor of the outcome variable Y (see also, Raudenbush, 1997; VanderWeele, 2008). Additionally, some of the statistical models that take into account that the empirical cluster means of the unit-covariate are only fallible measures of the values of the between-component Z_b require including the unit-covariate Z in the model (see, Asparouhov & Muthén, 2006; Lüdtke et al., 2008; B. O. Muthén,

1998-2004).

Full Adjustment Model: Alternative Parametrization

In the derivation of the full adjustment model for multilevel designs with treatment assignment at the cluster-level, the regression $E(Y \mid X, Z, V, Z_b)$ was parametrized using the within-component Z_w; an alternative parametrization uses the raw scores of the unit-covariate Z. The two parametrizations of $E(Y \mid X, Z, V, Z_b)$ give identical information about the conditional expectation of the outcome variable Y given the covariates on the between- and the within-level (Enders & Tofighi, 2007; Kreft et al., 1995) since the within-component Z_w is defined as the difference between the raw scores of the unit-covariate Z and the between-component Z_b. However, the meaning of the regression parameters and the resulting non-linear constraint differ. Since the implementation of the full adjustment model as multigroup multilevel latent variable model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007), that we will study in the simulation study in Section 5.2, requires the inclusion of the raw scores of the unit-covariate Z, we will briefly study the derivation of the identifier for the ACE_{10} for the alternative parametrization. The actual identification of the ACE_{10} in the multigroup multilevel model in `Mplus` is further derived in Appendix A.2.

If the regression $E(Y \mid X, Z, V, Z_b)$ is linear in the treatment variable X, the unit-covariate Z, the cluster-covariate V, the between-component of the unit-covariate Z_b and their products, it can also be written as:

$$\begin{aligned} E(Y \mid X, Z, V, Z_b) = {} & \gamma^*_{00} + \gamma^*_{01} Z_b + \gamma^*_{02} V + \gamma^*_{03} Z_b \cdot V \\ & + \left[\gamma^*_{04} + \gamma^*_{05} Z_b + \gamma^*_{06} V + \gamma^*_{07} Z_b \cdot V\right] \cdot Z \\ & + \left[\gamma^*_{10} + \gamma^*_{11} Z_b + \gamma^*_{12} V + \gamma^*_{13} Z_b \cdot V \right. \\ & \left. + \left[\gamma^*_{14} + \gamma^*_{15} Z_b + \gamma^*_{16} V + \gamma^*_{17} Z_b \cdot V\right] \cdot Z\right] \cdot X. \end{aligned} \tag{5.26}$$

The two conditional regressions $E_{X=0}(Y \mid Z, V, Z_b)$ and $E_{X=1}(Y \mid Z, V, Z_b)$ have the fol-

lowing form

$$\begin{aligned} E_{X=0}(Y \mid Z, V, Z_b) &= \gamma^*_{00} + \gamma^*_{01} Z_b + \gamma^*_{02} V + \gamma^*_{03} Z_b \cdot V \\ &\quad + [\gamma^*_{04} + \gamma^*_{05} Z_b + \gamma^*_{06} V + \gamma^*_{07} Z_b \cdot V] \cdot Z, \end{aligned} \tag{5.27}$$

$$\begin{aligned} E_{X=1}(Y \mid Z, V, Z_b) &= \gamma^*_{00} + \gamma^*_{10} + (\gamma^*_{01} + \gamma^*_{11}) Z_b \\ &\quad + (\gamma^*_{02} + \gamma^*_{12}) V + (\gamma^*_{03} + \gamma^*_{13}) Z_b \cdot V \\ &\quad + [\gamma^*_{04} + \gamma^*_{14} + (\gamma^*_{05} + \gamma^*_{15}) Z_b \\ &\quad + (\gamma^*_{06} + \gamma^*_{16}) V + (\gamma^*_{07} + \gamma^*_{17}) Z_b \cdot V] \cdot Z. \end{aligned} \tag{5.28}$$

The conditional effect function $g^*_1(Z, V, Z_b)$ is obtained by subtracting the extension of Equation (5.27) from the extension of Equation (5.28):

$$\begin{aligned} g^*_1(Z, V, Z_b) &= E^\circ_{X=1}(Y \mid Z, V, Z_b) - E^\circ_{X=0}(Y \mid Z, V, Z_b) \\ &= \gamma^*_{10} + \gamma^*_{11} Z_b + \gamma^*_{12} V + \gamma^*_{13} Z_b \cdot V \\ &\quad + [\gamma^*_{14} + \gamma^*_{15} Z_b + \gamma^*_{16} V + \gamma^*_{17} Z_b \cdot V] \cdot Z. \end{aligned} \tag{5.29}$$

The expected value of the conditional effect function $g^*_1(Z, V, Z_b)$ identifies the average causal effect ACE_{10}

$$\begin{aligned} E[g^*_1(Z, Z_b, V)] &= \gamma^*_{10} + \gamma^*_{11} E(Z) + \gamma^*_{12} E(V) + \gamma^*_{13} E(Z_b \cdot V) + \\ &\quad + \gamma^*_{14} E(Z) + \gamma^*_{15} E(Z \cdot Z_b) + \gamma^*_{16} E(Z \cdot V) + \gamma^*_{17} E(Z \cdot Z_b \cdot V). \end{aligned} \tag{5.30}$$

In contrast to Equation (5.23), Equation (5.30) cannot be further simplified: The identifier of the ACE_{jk} for the parametrization of $E(Y \mid X, Z, V, Z_b)$ with the raw scores of the unit-covariate Z includes the expected values of the products of the unit-covariate Z with the cluster-covariate V and the between-component Z_b. This inclusion is necessary to account for the shift of meaning in the parameters between the two parametrizations: In conventional multilevel terminology, the regression parameters associated with the between-component Z_b are the *between*-effects of Z if the within-component Z_w is used to parametrize the regression. If the raw score of the unit-covariate Z are used instead, these regression parameters represent the *contextual* influences of Z that are defined as the difference between the *between*- and the *within*-effect of the unit-covariate (see also Section 2.3 for an introduction of within- and between-effect in multilevel regres-

sions). The regression parameters associated with Z_w retain their meaning in the two parametrizations (Enders & Tofighi, 2007; Kreft et al., 1995). Even though the two parametrizations lead to identical *ACE*-estimators, parametrizing $E(Y \mid X, Z, V, Z_b)$ with the within-cluster residual Z_w is preferred for models with interactions because the coefficients are easier interpretable as effects on different hierarchical levels and model estimation is more stable because the predictor Z_w is regressively independent of Z_b (Enders & Tofighi, 2007). The multigroup multilevel latent variable model in `Mplus` 5.0, however, requires the use of the raw scores of the unit-covariate Z (L. K. Muthén & Muthén, 1998-2007) and further derivations to obtain the expected values of the product variables (see Appendix A.2 for further details).

5.2 Simulation Study

In the following section, we describe a large simulation study that compared the performance of several statistical implementations of the adjustment models for non-randomized multilevel designs with treatment assignment at the cluster-level with respect to their performance in finite samples. In this simulation, we tested various statistical models with an average causal effect of zero in the population to establish their adequacy under the null hypothesis. In order to simplify the simulation, we only considered a design with a unit-covariate Z and did not include a cluster-covariate V. However, the unit-covariate Z influenced the treatment assignment probabilities through its between-component Z_b and the outcome variable Y independently through both its between-component Z_b and its within-component Z_w.

The section is structured as follows: We first introduce the data generation procedure and assumptions made in repeating the single-unit trial. Next, we introduce the statistical models and research questions to be addressed in the simulation study. We then describe the design of the simulation study and report the results: The *ACE*-estimators were studied with respect to their bias and their relative efficiency, the standard errors of the *ACE*-estimators were studied with respect to their bias and the empirical type-1-error rate in tests of the null hypothesis.

5.2.1 Data Generation

The generalized ANCOVA was developed in the previous section with reference to the single-unit trial and the causality space introduced in Chapter 2 (see also, Steyer et al., 2009). Inferences from finite samples that are the focus of the simulation study require independent repetitions of a single-unit trial that are stable with respect to causal parameters and distributions and a statistical model to estimate the average causal effect and its standard error (see also the discussions in Chapter 3). In this section, we describe the central components of the repeated single-unit trial in the simulation study and discuss the resulting properties of the sampling model. A detailed description of the implementation of the data generation in R (R Development Core Team, 2008) is given in Appendix C.2.

In line with the theoretical concepts and data generation procedures in other simulation studies of the analysis of average causal effects with the generalized ANCOVA for singlelevel designs (e.g., Kröhne, 2009), data was generated by considering the regressions of the true-outcome variable τ_0 and the true-effect variable δ_{10} on the covariates Z and Z_b and their respective residuals ε_0 and ε_{10}:

$$\tau_0 = E(\tau_0 \mid Z, Z_b) + \varepsilon_0, \tag{5.31}$$

$$\delta_{10} = E(\delta_{10} \mid Z, Z_b) + \varepsilon_{10}, \tag{5.32}$$

where $E(\delta_{10} \mid Z, Z_b)$ is the conditional effect function $CCE_{10;Z,Z_b}$. The presence of the residuals ε_0 and ε_{10} violated conditional homogeneity as introduced in Section 2.6.2: The true-outcome variable τ_0 in the control condition and the true-effect variable δ_{10} were not constant given the covariates Z and Z_b. More specifically, the residuals ε_0 and ε_{10} were assumed to be further decomposable into cluster-specific components $r_{j;C}$ and unit-specific components $\upsilon_{j;U}$ to represent the effects of the cluster variable C and the unit variable U

$$\varepsilon_0 = r_{0;C} + \upsilon_{0;U}, \tag{5.33}$$

$$\varepsilon_{10} = r_{10;C} + \upsilon_{10;U}. \tag{5.34}$$

The unit-specific components $\upsilon_{0;U}$ and $\upsilon_{10;U}$ are defined as the residuals of the regressions $E(\tau_0 \mid Z, Z_b, C)$ and $E(\delta_{10} \mid Z, Z_b, C)$ respectively. As residuals, their expected value

is equal to zero and their regression on the regressors Z, Z_b and C is zero. Furthermore, their covariance $Cov\,(\upsilon_{0;U}, \upsilon_{10;U})$ was fixed to zero in the simulation. The residuals $r_{0;C}$ and $r_{10;C}$ accounted for the multilevel structure of the repeated single-unit trial: They captured the residual influence of the cluster variable C on the true-outcome variable τ_0 and the true-effect variable δ_{10}. The residuals $r_{0;C}$ and $r_{10;C}$ are defined as residuals of $E\,[E(\tau_0 \mid Z, Z_b, C) \mid Z, Z_b]$ and $E\,[E(\delta_{10} \mid Z, Z_b, C) \mid Z, Z_b]$ respectively. Hence, their expected value is equal to zero and their regression on the regressors Z and Z_b is zero. Also, their covariance with the unit-specific residuals $\upsilon_{0;U}$ and $\upsilon_{10;U}$ is zero. Additionally, their covariance $Cov\,(r_{0;C}, r_{10;C})$ was equal to zero in the data generation procedure. If the variance of the residual ε_{10} is larger than zero, there will be residual variance heterogeneity between the treatment groups. Depending on whether this heterogeneity is due to $Var(r_{10;C}) > 0$ or $Var(\upsilon_{10;U}) > 0$, the heterogeneity is located at the unit- or at the cluster-level or at both levels.

The decomposition of the residuals ε_0 and ε_{10} yielded the following residual intraclass correlation coefficients *rICC*s in line with the general definition in Equation (2.11):

$$rICC(\tau_0 \mid Z, Z_b) = \frac{Var\,[E(\tau_0 \mid Z, Z_b, C) - E(\tau_0 \mid Z, Z_b)]}{Var\,[\tau_0 - E(\tau_0 \mid Z, Z_b)]} = \frac{Var(r_{0;C})}{Var(r_{0;C}) + Var(\upsilon_{0;U})}, \tag{5.35}$$

$$rICC(\delta_{10} \mid Z, Z_b) = \frac{Var\,[E(\delta_{10} \mid Z, Z_b, C) - E(\delta_{10} \mid Z, Z_b)]}{Var\,[\delta_{10} - E(\delta_{10} \mid Z, Z_b)]} = \frac{Var(r_{10;C})}{Var(r_{10;C}) + Var(\upsilon_{10;U})}. \tag{5.36}$$

If Equations (5.35) and (5.36) are both equal to zero, i.e., if the variances of the cluster-specific components $r_{0;C}$ and $r_{10;C}$ are both equal to zero, the estimated distribution of the parameter estimates from a conventional singlelevel regression model will be correct. If these variances are different from zero, the standard errors and covariances of parameter estimates will be underestimated, if the corresponding variance components of the outcome variable Y are not included in the statistical model (Hedges, 2007a; Moerbeek et al., 2000, 2001; Raudenbush, 1997; Snijders & Bosker, 1999).

The regressions $E(\tau_0|Z, Z_b)$ and $E(\delta_{10}|Z, Z_b)$ were parametrized as linear functions of the within-component Z_w, the between-component Z_b and their product using the same

labels for the regression coefficients as in Equation (5.18):

$$E(\tau_0 \mid Z, Z_b) = \gamma_{00} + \gamma_{01} \cdot Z_b + \gamma_{04} \cdot Z_w + \gamma_{05} \cdot Z_b \cdot Z_w, \tag{5.37}$$

$$E(\delta_{10} \mid Z, Z_b) = \gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w. \tag{5.38}$$

Both, Equations (5.37) and (5.38), used the true values z_b of the between-component Z_b, the (C=c)-conditional expected values $E(Z|C{=}c)$, and the residual of this regression, the within-component Z_w. However, the true values of Z_b were not available in the sample and had to be approximated by the cluster-means of Z and the empirical deviations from the cluster-means. If the regression weights of the between- and the within-component in Equations (5.37) and (5.38) are equal, i.e., if $\gamma_{01} = \gamma_{04}$ and $\gamma_{11} = \gamma_{14}$, and if their are no interactions between Z_b and Z_w, i.e., $\gamma_{05} = \gamma_{15} = 0$, or if the $ICC(Z)$=0, the multilevel decomposition of the unit-covariate does not have to be taken into account explicitly [see Equation (2.13)]. If at least one of these conditions is not fulfilled, a statistical model that only includes the unit-covariate Z as predictor will be misspecified and lead to a biased estimator of the ACE. The cluster variable C did not interact with the within-component Z_w, i.e., the parameters γ_{04} and γ_{14} were constant across clusters. The average causal effect is the expected value of the true-effect variable δ_{10} given in Equation (5.32) using the decomposition of the residual ε_{10} in Equation (5.34) and the parametrization in Equation (5.38):

$$ACE_{10} = E(\gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w + r_{10;C} + \upsilon_{10;U}), \tag{5.39}$$

$$= \gamma_{10} + \gamma_{11} \cdot E(Z). \tag{5.40}$$

Treatment assignment probabilities were determined by a logistic function that described the probability of clusters being assigned to each treatment condition as a function of the between-component Z_b:

$$P(X{=}1 \mid Z_b) = \frac{exp\,(g_0 + g_1 \cdot [Z_b - E(Z)])}{1 + exp\,(g_0 + g_1 \cdot [Z_b - E(Z)])}. \tag{5.41}$$

This function guaranteed independence of the treatment variable X and the confounder σ-algebra $\mathfrak{C}_X$ given the between-component Z_b (i.e., $X \perp \mathfrak{C}_X|Z_b$) as a sufficient condition for unbiasedness and unconfoundedness of $E(Y \mid X, Z_b)$. If the parameter g_1 is equal to zero, the treatment assignment probabilities do not depend on the between-component

Z_b and the data generation would represent a cluster-randomized design with equal treatment probabilities for each cluster.

Each data set for the simulation study was generated by repeating the single-unit trial as described above with further assumptions about the distributions and parameters involved. By sampling from the unconditional distribution of the unit-covariate Z in each repetition of the single-unit trial to represent quasi-experimental designs (see Kröhne, 2009), the realized values of Z varied from sample to sample. Hence, the unit-covariate Z and by implication the between-component Z_b and the within-component Z_w are stochastic predictors (Chen, 2006; Nagengast, 2006; Sampson, 1974; Shieh, 2006). In a similar vein, the use of the probabilistic assignment function, given in Equation (5.41), yielded samples that varied in the treatment group sizes depending on the realized values of the covariate and the actual assignment of units to treatment conditions in the repeated single-unit trials. Hence, the treatment variable X also was a stochastic predictor (Nagengast, 2006). A detailed description of the parameters that were varied or kept constant within the simulation design is given in Section 5.2.3. The implementation of the data generation in `R` is explained and corresponding parameter values of the data generation procedure are given in Appendix C.2.

Summarizing, the data generation routine resulted in four special properties of the sampling model that were represented to a different degree by the statistical models in the simulation study:

1. Due to repeated sampling from the unconditional distribution of the unit-covariate Z, its realized values varied from sample to sample. Thus, the unit-covariate Z and by implication the between-component Z_b and the within-component Z_w were stochastic predictors (Chen, 2006; Sampson, 1974; Shieh, 2006). The same logic applied to the realizations of the treatment variable X.

2. The conditional causal effect function $CCE_{10\,;\,Z,\,Z_b}$ varied independently with both the between-component Z_b and with the within-component Z_w. The treatment assignment probabilities, however, were only influenced by the between-component Z_b. Thus, Z_b is a relevant covariate to be considered in adjustment models, if $ICC(Z) > 0$ [see Equation (2.13)].

3. The regressions $E(\tau_0 \mid Z, Z_b)$ and $E(\delta_{10} \mid Z, Z_b)$ were specified using the actual values of the regression $E(Z \mid C{=}c)$. These values are only approximated by the

empirical cluster means of Z that are fallible measures of $E(Z \mid C)$ in samples (Asparouhov & Muthén, 2006; Lüdtke et al., 2008). The average reliability of the cluster means is a function the average cluster sizes and determines whether the estimated regression coefficients associated with the between-component Z_b will be biased [see Equation (2.14)].

4. Both $E(\tau_0 \mid Z, Z_b)$ and $E(\delta_{10} \mid Z, Z_b)$ were allowed to have residual intraclass correlation coefficients ($rICC$) larger than zero, reflecting systematic influences of the cluster-variable C on τ_0 and δ_{10} after taking the effects of the unit-covariate Z and the between-component Z_b into account. In addition, the variances of all residuals were not restricted to be equal, potentially resulting in residual variance heterogeneity of the outcome variable Y between treatment groups.

In the following section, we will discuss several statistical models and their abilities to deal with the properties of the data generation procedure.

5.2.2 Research Questions and Statistical Methods

The simulation study investigated the finite-sample performance of several statistical implementations of the generalized ANCOVA under the null hypothesis of no average causal effect ($H_0 : ACE_{10} = 0$). In line with the peculiarities of the data generation procedure, we tested the robustness of various statistical methods against violations of their assumptions and addressed the following research questions and hypotheses:

1. The decomposition of the unit-covariate Z into the between-component Z_b and the within-component Z_w has to be accounted for in the statistical analysis of conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the cluster-level, if Z_b and Z_w independently influence the conditional causal effect function $CCE_{jk;Z,Z_b}$. Neglecting the decomposition and specifying a naive adjustment model that uses only the unit-covariate Z as predictor leads to a considerable bias in the ACE-estimator (Snijders & Bosker, 1999).

2. Even if the model is correctly specified in its fixed part and takes the decomposition of Z into the between-component Z_b and the within-component Z_w into account, residual effects of the cluster variable C need to be modeled by including variance components for $Var(r_{0;C})$ and $Var(r_{10;C})$. Statistical models that do

not include the additional variance components will lead to standard errors that underestimate the variability of the *ACE*-estimates.

3. Statistical models of contextual effects have to account for the fact that the empirical cluster-means of the unit-covariate Z are fallible measures of the corresponding values of the regression $E(Z \mid C)$. The resulting *ACE*-estimators will be biased [see Equation (2.14)], if the measurement error is not explicitly accounted for by using the multilevel latent variable model (Lüdtke et al., 2008) or other adjustment procedures correcting for the unreliability of the cluster means (Croon & van Veldhoven, 2007).

4. The *ACE*-estimators obtained from the full and the simple adjustment model as introduced in Section 5.1.2 will be unbiased and identical (Snijders & Bosker, 1999; VanderWeele, 2008), if they are obtained with the same statistical method. However, the estimator from the full adjustment model will be more efficient.

5. Statistical models that include the appropriate variance components for $Var(r_{0;C})$ and $Var(r_{10;C})$, but do not treat the unit-covariate Z as a stochastic predictor — implicitly assuming that the empirical mean of Z is equal to the expected value $E(Z)$ and constant over replications — will underestimate the variability of the corresponding *ACE*-estimates, especially when there are strong interactions between the treatment variable X and the covariates (Kröhne, 2009; Nagengast, 2006). Methods that take the stochasticity of Z explicitly into account by including $E(Z)$ as a model parameter will yield accurate standard errors of the *ACE*-estimator in all conditions.

Statistical details of the models and the implementation of the generalized ANCOVA are given in Appendix B. An overview of the properties of the statistical models is given in Table 5.1 with respect to the properties of the repeated single-unit trial and the resulting research questions. Specifically, the following implementations of the generalized ANCOVA were compared in the simulation study:

- The *naive singlelevel model* implementation of the generalized ANCOVA model in `lace` (Partchev, 2007), using only the unit-covariate Z as a predictor in separate group-specific structural equation models and neglecting the multilevel structure of the unit-covariate Z by not further decomposing Z into the between-component

Table 5.1: Properties of the statistical models to implement the generalized ANCOVA for designs with treatment assignment at the cluster-level

Method	Multilevel effects of Z	Stochastic predictors	Variance components	Latent variable Z_b
`lace`: Naive model	-	*x*	-	-
`lace`: Full model	*x*	*x*	-	-
`nlme`: Simple model	*x*	-	*x*	-
`nlme`: Full model	*x*	-	*x*	-
`Mplus`: Full model (singlegroup)	*x*	*x*	*x*	-
`Mplus`: Full model (multigroup)	*x*	*x*	*x*	*x*
Croon: Simple Model (GLH)	*x*	-	*(x)*	*x*
Croon: Simple Model (`lace`)	*x*	*x*	*(x)*	*x*

Z_b and the within-component Z_w, [see Equation (B.5) in Appendix B.1 for the model specification];

- the *full adjustment model* in `lace`, separately modeling the between-component Z_b and the within-component Z_w, using the cluster-means of the unit-covariate Z, the cluster-mean centered values of Z and their products as predictors in group-specific structural equation models, but not including variance components for the residuals $r_{0;C}$ and $r_{10;C}$, [see Equation (B.6) in Appendix B.1 for the model specification];

- the *simple adjustment model* in `nlme` (Pinheiro et al., 2008) modeling the between-component Z_b, using the cluster-means of the unit-covariate Z, the treatment indicator and their product as predictor, and modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by including a random intercept (assuming homogeneity of its variance in treatment and control group), but obtaining standard errors and significance tests of the *ACE* with the general linear hypothesis, thereby treating Z as a fixed predictor [see Equation (B.12) in Appendix B.2 for the model specification];

- the *full adjustment model* in `nlme` (Pinheiro et al., 2008) separately modeling the between-component Z_b and the within-component Z_w, using the cluster-means of the unit-covariate Z and the cluster-mean centered values of Z as predictors, and

modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by a random intercept (assuming homogeneity of its variance in treatment and control group), but obtaining standard errors and significance tests of the *ACE* with the general linear hypothesis, thereby treating Z as a fixed predictor [see Equation (B.13) in Appendix B.2 for the model specification];

- the *full adjustment model* in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) implemented as a singlegroup multilevel model, separately modeling the between-component Z_b and the within-component Z_w, using the cluster-means of the unit-covariate Z, the cluster-mean centered values of Z, the treatment indicator and their products as predictors, modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by including a random intercept (assuming homogeneity of its variance in treatment and control group), and estimating the expected value of the unit-covariate Z as a model parameter and thereby taking the stochasticity of the unit-covariate Z explicitly into account in the estimation of the *ACE* and its standard error [see Equation (B.30) in Appendix B.3 for the model specification];

- the *full adjustment model* in `Mplus` 5.0 implemented as a multigroup multilevel latent variable model, separately modeling the between-component Z_b and the within-component Z_w as latent variables, using them as predictors in treatment group specific regression models, modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$ by including random intercepts (whose variances varied between treatment groups) and estimating the expected value of the unit-covariate Z as a model parameter and thereby taking the stochasticity of the unit-covariate Z explicitly into account in the estimation of the *ACE* and its standard error, [see Appendix A.2 for the derivation of the non-linear constraint to identify the *ACE* and Equation (B.36) in Appendix B.3 for the model specification];

- the *simple adjustment model* specified with the two-step adjustment procedure of Croon and van Veldhoven (2007) modeling the between-component Z_b with the reliability-corrected cluster means of Z, the treatment indicator and their product as predictors in a general linear model with the reliability corrected cluster means of Y as outcomes (thereby implicitly modeling the variance components for the residuals $r_{0;C}$ and $r_{10;C}$), using either the general linear hypothesis to obtain standard errors and significance tests of the *ACE* or alternatively using `lace`

to estimate the *ACE* and its standard error [see Equation (B.46) for the model specification using the GLH and Equation B.47 for the model specification with `lace`].

There were several reasons, why the research questions with respect to the statistical methods required a simulation study and could not be directly addressed by analytical derivations:

1. The distributional theory for the statistical models only holds asymptotically (see Appendix B for further details). Their performance and properties in finite samples under realistic conditions determines their usefulness for applications.

2. The standard errors of the *ACE*-estimator in `lace` and `Mplus` obtained with the multivariate delta-method are — even asymptotically — first-order Taylor-series approximations (Rao, 1973; Raykov & Marcoulides, 2004). Thus, simulation studies are called for to analyze their properties and appropriateness in finite samples (see also MacKinnon, 2008).

3. Finally, some of the statistical methods do not account for *all* peculiarities of the data generation procedure while those methods that do, potentially require larger sample sizes. The potential trade-off of sample size requirements and robustness to violations is thus important to be considered.

5.2.3 Design

Again, the simulation was implemented in `R` (R Development Core Team, 2008) using `SimRobot` (Kröhne, 2007) to manage and distribute the simulation conditions on a cluster of 40 workstations. All simulated data sets were generated with the data generation routine for designs with treatment assignment at the cluster-level described in Appendix C.2. In this section, the parameters that were varied — the independent variables of the simulation design — and the parameters that were held constant over simulation conditions are described in detail.

Independent Variables

The following parameters of the data generation routine were varied in a five-factorial fully-crossed simulation design with 1000 replications per cell: (1) the number of clus-

Table 5.2: Factors of the simulation design for designs with treatment assignment at the cluster-level

Factor	Measure	Values
Number of clusters		20, 50, 100, 200
Average cluster sizes	$\overline{n}_c$	5, 10, 25, 50
Intra-class-correlation of Z	$ICC(Z)$	0.05, 0.1, 0.2, 0.3
Dependency of X and Z_b	$\sqrt{R^2}$	0, 0.15, 0.3, 0.5
Effect size of cluster-level interaction	$d(\gamma_{11})$	0, 0.025, 0.05, 0.1, 0.2, 0.3

ters, (2) the average number of sampled units within each cluster, (3) the intraclass correlation coefficient (ICC) of the unit-covariate Z, (4) the dependency between the treatment variable X and the between-component Z_b and (5) the effect size of the interaction between the treatment variable X and the between-component Z_b. We will describe the values of the independent variables in the simulation design and outline the motivation behind these choices. An overview of the design factors is given in Table 5.2. The corresponding parameters of the data generation routine are given in Table C.3 in Appendix C.2.

Number of Clusters. The total number of clusters was either $20, 50, 100$ or 200. 20 clusters were chosen as a sensible lower bound for the total number of clusters, since cost-effective designs for cluster-randomized trials use a relatively small number of clusters with medium to large number of units per cluster (Feng et al., 2001; Raudenbush, 1997). Additionally, it was deemed unlikely that a smaller sample of clusters would still result in well-behaved statistical models. Previous simulation studies of the hierarchical linear model (e.g., Browne & Draper, 2000; Maas & Hox, 2005) have usually considered larger sample sizes at the cluster-level and have reported adequate performance of asymptotic estimators for models with random slopes in terms of estimation bias and coverage for samples starting at 50 clusters. Lüdtke et al. (2008) reported acceptable behavior of the multilevel latent variable model in `Mplus` for a random intercept model starting at a sample size of 100 clusters. 200 clusters were chosen as an upper bound for the number of clusters that still might be plausibly obtained for between-group multilevel designs with treatment allocation at the cluster-level in empirical applications (Murray, 1998). Unsatisfactory performance of the adjustment

methods at this number of clusters would render them unsuitable for applications.

Average Cluster Sizes. The following average number of units per cluster $\overline{n}_c$ were considered: $5, 10, 25, 50$. Actual cluster sizes varied by 10% around the average size to represent naturally occurring variations in cluster sizes in designs with pre-existing clusters. The total sample size for each replication was fixed to the product of the number of clusters and the average cluster size. The average cluster sizes were chosen to represent a wide range of realistic values in applications. Cluster sizes of 5 can be expected to be encountered in small group research, e.g., working groups in organizational studies (Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Average cluster sizes of 10 are relevant in evaluation studies of group interventions in psychology and medicine (e.g., Helgeson et al., 1999). In educational studies, class sizes of about 25 students can be expected, at least in the context of the German educational system (Lüdtke et al., 2008). Finally, cluster sizes as large (or even larger) than 50 are the norm in cluster-randomized trials in medical interventions (Feng et al., 2001) or in neighborhood research (Oakes, 2004; VanderWeele, 2008).

Intraclass Correlation of the Unit-Covariate. The following values of the intraclass correlations of the unit-covariate Z [$ICC(Z)$] were considered: $0.05, 0.1, 0.2$ and 0.3. They were chosen to cover the range of typical intraclass correlation coefficients found in medical and educational research in accordance with the reviews by Gulliford et al. (1999), Hedges and Hedberg (2007) and Schochet (2008). An $ICC(Z)$ of 0.3 is higher than most of the empirical ICC-values reported in these studies. In educational studies, values between 0.1 and 0.2 are fairly common (Hedges & Hedberg, 2007; Schochet, 2008). In medical research, values of around 0.05 are normal, when reasonable cluster sizes are considered (Gulliford et al., 1999). An $ICC(Z)$ of zero, as expected in designs with random assignment of units to clusters, was excluded, since such designs are seldom in applications (Murray, 1998). The $ICC(Z)$ was manipulated by varying the variance of Z_b, while holding the variance of Z_w constant at a value of 1. The exact values of the corresponding variance parameters can be found in Appendix C.2.

Dependency of X and Z_b. The dependency of the treatment variable X and the between-component Z_b was varied to represent different degrees of confounding. For this purpose, the parameter g_1 of the logistic assignment function [see Equation (5.41)]

was chosen to keep the square root of the coefficient of determination of the logistic assignment function $\sqrt{R^2}$ (Nagelkerke, 1991) constant across different values of $ICC(Z)$. Specifically, the following values were studied: $\sqrt{R^2} = 0, 0.15, 0.3, 0.5$. When $\sqrt{R^2} = 0$, the treatment assignment probabilities did not depend on the values of the between-component Z_b and the design was actually a cluster-randomized experiment. The parameters of the logistic assignment function resulted in equal unconditional probabilities for the treatment and the control group in all conditions of the simulation design [$P(X{=}0){=}P(X{=}1){=}0.5$], no matter how strong the dependencies between the treatment variable X and between-component Z_b were. The exact values of g_1 for different ICC-values of Z were obtained in an exploratory simulation study and are exact up to the third decimal. The parameter values are given in Table C.3 in Appendix C.2.

Effect Size of Cluster-Level Interaction. Finally, the effect size of the interaction between X and Z_b was varied as a factor in the simulation design. In order to define an effect size measure for the interaction that was independent of the strength of association between X and Z_b, the effect size was measured by the proportion of the sum of the variance of the true-outcome variable τ_1 in the treatment condition and the residual variance σ^2_Y of the outcome variable Y that was due to the interaction γ_{11}:

$$d(\gamma_{11}) = \frac{Var(\gamma_{11} \cdot Z_b)}{Var(\tau_1) + \sigma^2_Y} = \frac{\gamma^2_{11} \cdot Var(Z_b)}{Var(\tau_0 + \delta_{10}) + \sigma^2_Y}. \tag{5.42}$$

The regression weight γ_{11} is the regression weight of the product variable of X and Z_b in Equation (5.18). $d(\gamma_{11})$ was set to values of $0, 0.025, 0.05, 0.1, 0.2$ and 0.3 to represent realistic to extreme effect sizes of interactions in applications. Since the variance of the between-component Z_b varied with different values of $ICC(Z)$, the corresponding parameter values for γ_{11} had to be varied accordingly to keep $d(\gamma_{11})$ constant for different conditions of $ICC(Z)$. These values were obtained analytically using `YACAS` (Goedman et al., 2007). The corresponding parameters of the data generation function are documented in Table C.3. In all cases, the regression weight γ_{11} of the product of X and Z_b was positive or equal to zero, while the regression weight γ_{14} of the product of X and Z_w, that was not varied in the simulation, was negative.

Constant Parameters

Average Causal Effect. The goal of the simulation was to study the performance of the different implementations of the adjustment model under the null hypothesis of no average causal effect ($H_0 : ACE_{10} = 0$). The model parameters were chosen to guarantee that the average causal effect ACE_{10} was fixed at a value of zero in all experimental conditions. The expected value of the unit-covariate $E(Z)$ was set to 1. Since, the regression weight γ_{11} of the product of X and Z_b varied with the effect size of the interaction and different values of $ICC(Z)$, the intercept γ_{10} of the conditional effect function $CCE_{10;Z,Z_b}$ was chosen accordingly to fix the average causal effect to zero in all cells of the simulation design. An overview of all fixed regression weights of the simulation is given in Table C.4 in Appendix C.2.

Variance Parameters. The following variance parameters were constant over all experimental conditions: $\sigma^2_Y = 2$, $\sigma^2_{Z_w} = 1$, $\sigma^2_{\upsilon_{0;U}} = 2.25$, $\sigma^2_{\upsilon_{10;U}} = 1.25$, $\sigma^2_{r_{0;C}} = 0.75$ and $\sigma^2_{r_{10;C}} = 0.25$. These settings resulted in a moderate heterogeneity of unit- and cluster-level variances of the outcome variable Y. However, the $(X{=}j)$-conditional residual intraclass correlation coefficients $rICC_{X=j}(Y \mid Z, Z_b)$ of the outcome variable Y were almost equal in treatment and control group (0.15 in the control group; 0.153 in the treatment group). The design effect (Kish, 1965) was larger than two in all sample size conditions (all residual VIFs > 2). Analyses that neglected the multilevel structure of the data would result in significantly underestimated standard errors of the model parameters (Hox, 2002; Maas & Hox, 2005). The average reliabilities of the cluster means as measures of the between-component Z_b ranged from 0.208 in the conditions with the smallest average cluster sizes and the smallest $ICC(Z)$ to 0.955 in the conditions with the largest average cluster sizes and the largest $ICC(Z)$. An overview of all fixed variance parameters in the simulation design is given in Table C.4 in Appendix C.2.

5.2.4 Results

In this section, we present the main results of the simulation study. We begin with reporting the convergence rates of the different methods, then give the bias of the ACE-estimators and its standard errors. Finally, we report the empirical type-1-error rates

Table 5.3: Average convergence rates: Full adjustment model implemented as singlegroup multilevel model in `Mplus`

	Average Cluster Sizes			
Number of Clusters	5	10	25	50
20	60.09%	72.32%	79.11%	81.70%
50	91.09%	96.86%	98.66%	98.99%
100	98.99%	99.89%	99.99%	99.99%
200	99.99%	100%	100%	100%

of the significance tests and briefly compare the mean squared errors of selected implementations. The dependent measures are introduced in Appendix D together with the cut-off criteria considered as boundaries for appropriate performance. At the end of the section, we summarize and review the results with respect to the research questions introduced above. We will only present selected results as tables and figures in the text. All results for the different methods and dependent measures are provided as graphics on the accompanying CD, as are the raw results for all simulation conditions (see Appendix E).

Convergence

Convergence problems hampered the performance of the singlegroup multilevel model in `Mplus`. These convergence problems were especially prevalent in conditions with 20 clusters, an effect that was only partially offset by larger average cluster sizes: An average of 81.70% of the models converged when the average cluster size was 50; only 60.09% of the models converged, when the average cluster size was 5. Some convergence problems were still present with samples of 50 clusters, where convergence rates ranged from 91.09% to 98.99%. Satisfactory average convergence rates were only obtained starting from samples of 100 clusters upwards, with all rates being close to or above 99% (see Table 5.3 for details). The other factors of the simulation design did not influence the convergence patterns. The convergence problems were not alleviated by including the true parameter values as starting values nor by loosening the convergence criteria or increasing the number of iterations.

A similar picture emerged for the `Mplus` multigroup multilevel latent variable model.

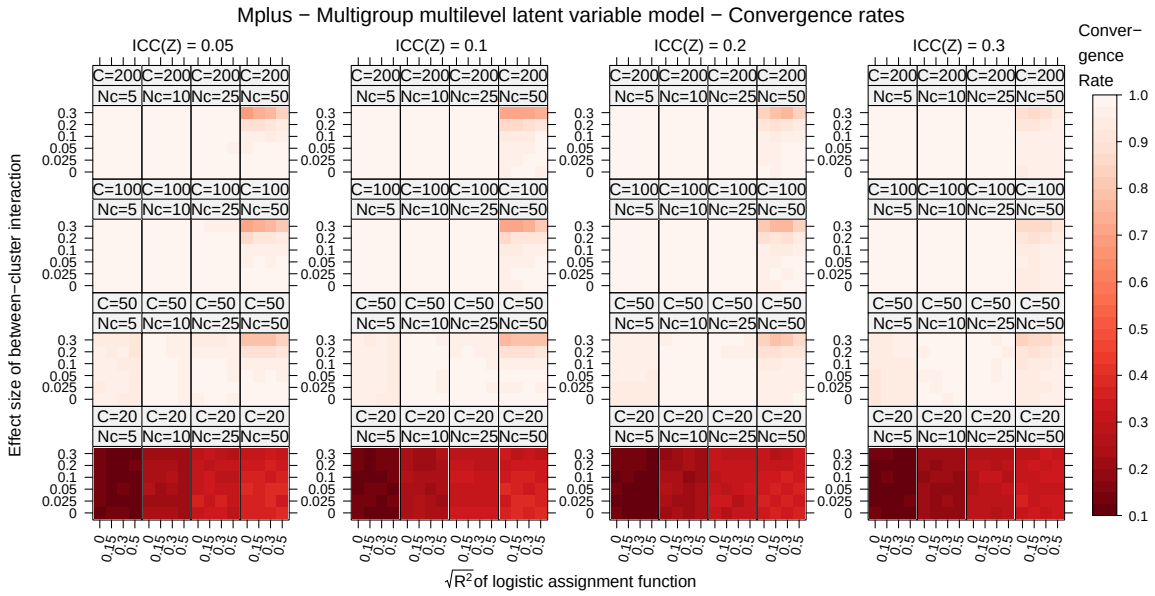

Figure 5.1: Convergence rates: Full adjustment model implemented as multigroup multilevel latent variable model in `Mplus`

At sample sizes of 20 clusters, average convergence rates ranged from 12.09% to 34.06%, again getting better with larger average cluster sizes. Surprisingly, the simulation conditions with an average cluster size of 50 and the largest effect size of the interaction at the cluster-level also had convergence problems: The average convergence rate in these cells was 80.14%. In the other cells of the design the average convergence rates were satisfactory (97.94%) (for details see Figure 5.1), although slightly worse in cells with large effect sizes of the interaction at the cluster-level and average cluster sizes of 50 and in cells with the smallest average cluster size ($\overline{n}_c = 5$), 50 clusters and an $ICC(Z) = 0.05$. Again, the convergence problems were not alleviated substantially by including the true parameter values as starting values nor by loosening the convergence criteria or increasing the number of iterations. An overview of the convergence rates is given in Figure 5.1.

All other methods did not exhibit significant convergence problems: Average convergence rates were well above 99% in all conditions, except for the combination of 20 clusters with an average cluster size of 5. Even in these extreme conditions, not a single average convergence rate was below 98.5%, indicating no serious convergence problems.

Bias of *ACE*-Estimator

In this section, we describe the bias in estimation of the average causal effect for the different implementations of the adjustment model. Our discussion will be organized as follows: We start with presenting the results for the naive adjustment model in `lace`, followed by a discussion of the models that used the empirical cluster-means and cluster-mean centered values of the unit-covariate Z as predictors. The *ACE*-estimators were biased considerably in both cases. The presentation concludes with the methods that provided unbiased estimators of the *ACE*: the multigroup multilevel latent variable model in `Mplus` and the adjustment procedure of Croon and van Veldhoven (2007). We will use the mean bias (*MB*) of the *ACE*-estimator as defined in Equation (D.1) to evaluate the different estimators. Following the recommendations by Boomsma and Hoogland (2001), an over- or underestimation of the *ACE* by 2.5% (corresponding to a *MB* between -0.025 and 0.025) was considered as threshold for unbiasedness of an estimator.

Unsuitable Methods. The naive adjustment model in `lace`, that only included the unit-covariate Z as predictor without taking the multilevel decomposition of Z into account, led to a highly biased estimator of the average causal effect. The mean bias averaged over all conditions of the simulation design was 0.08. A closer look revealed that the estimator was unbiased, if the between-component Z_b did not influence the treatment assignment, i.e., in cluster-randomized experiments. However, with stronger dependencies of Z_b and the treatment variable X, the *ACE*-estimator showed an increasingly large positive bias when an interaction on the cluster-level was present and an increasingly strong negative bias when there were no or only small interactions at the cluster level. The negative bias was magnified with larger $ICC(Z)$-values and the pattern of positive bias was shifted downwards, appearing only with small negative regression weights of the interaction or no interaction on the cluster-level. Average cluster size and the number of clusters did not influence the bias pattern considerably. The mean biases of the *ACE*-estimator for all conditions of the simulation design are given in Figure 5.2.

Methods that use Empirical Cluster Means as Predictors. The full adjustment model as implemented in `lace`, in `nlme` and as singlegroup multilevel model in `Mplus` with cluster means and cluster-mean centered values of the unit-covariate Z as predic-

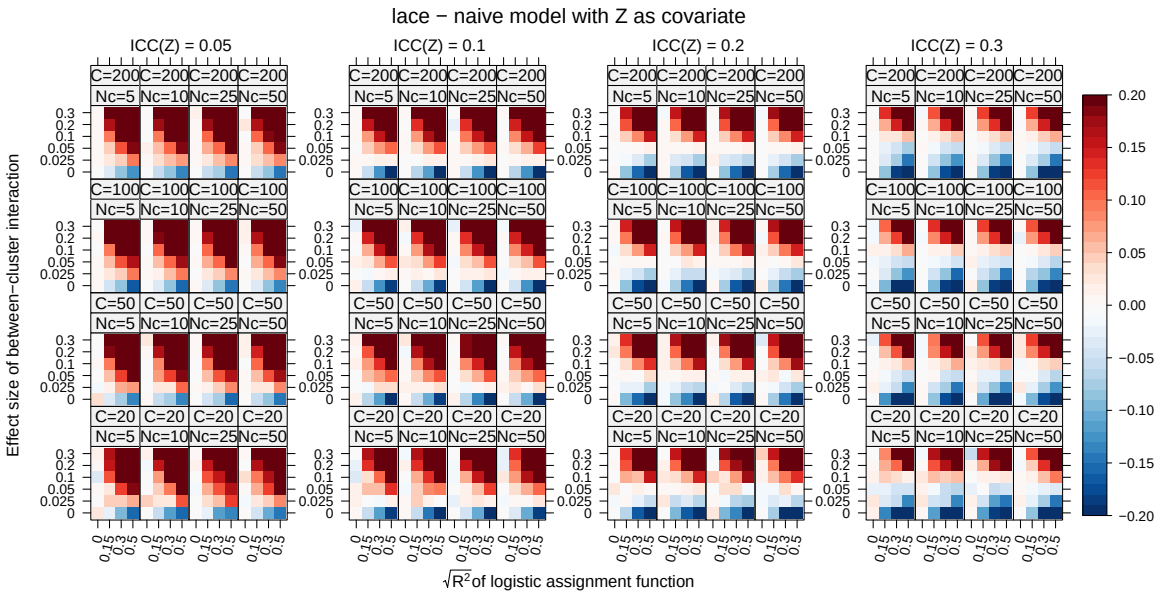

Figure 5.2: Mean bias of *ACE*-estimator: Naive adjustment model implemented in `lace`

tors and also the simple adjustment model specified in `nlme` gave identical results up to the third decimal as far as bias of the *ACE*-estimator was concerned: The mean biases were correlated almost perfectly (all $rs > 0.997$) over the conditions of the simulation design. Thus, the following presentation of the results will not differentiate any further between these methods.

In line with the analytical derivations by Lüdtke et al. (2008), the *ACE*-estimators obtained from these methods were strongly biased when the $ICC(Z)$ was small and the average cluster size was small [see Equation (2.14)]. The direction of the bias was influenced by the magnitude of the interaction effect on the cluster-level. The results indicated a positive bias with larger effect sizes of the interaction at the cluster-level $d(\gamma_{11})$: In these cases, the effect of the product variable $Z_w \cdot X$, the within-cluster interaction, differed most strongly from the effect of the product variable $Z_b \cdot X$, the between-cluster interactions, with the regression weight at the cluster-level γ_{11} being larger than the regression weight at the unit-level γ_{14}. When no interaction effect was present at the cluster-level, the *ACE*-estimator showed a negative bias. In these conditions, the difference between unit- and cluster-level regression effect was negative ($\gamma_{11} < \gamma_{14}$) and the *ACE* was underestimated. As expected from the analytical derivations by Lüdtke

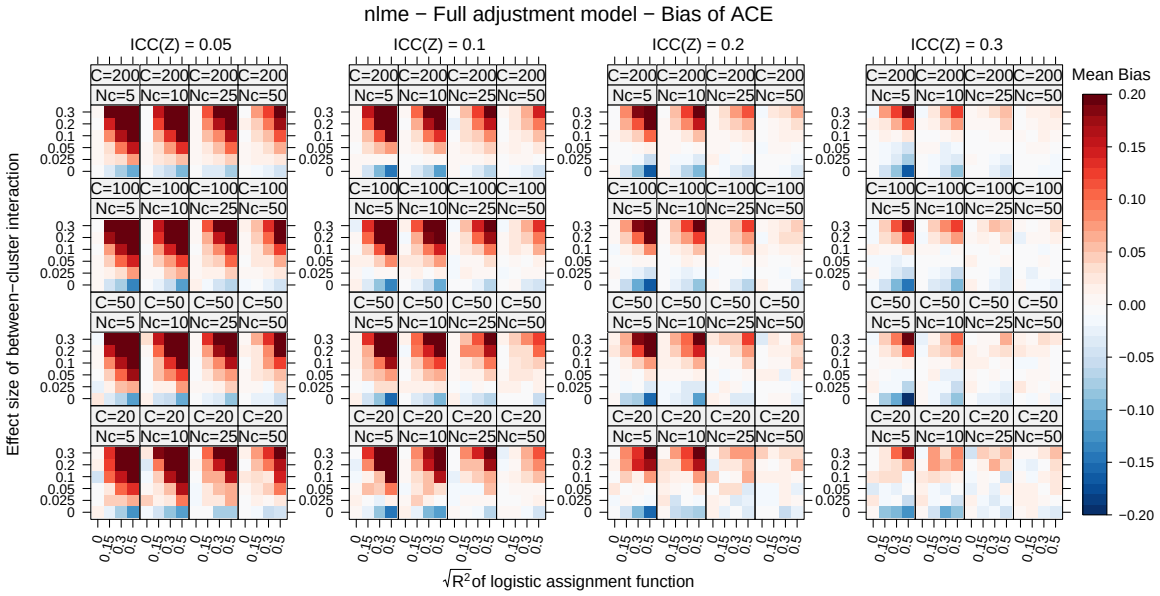

Figure 5.3: Mean bias of *ACE*-estimator: Methods that use cluster means and group-mean centered covariates as predictors

et al., the bias was significantly reduced with higher *ICC*s of the unit-covariate Z and with larger average cluster sizes. The pattern of over- and underestimation also shifted slightly with larger $ICC(Z)$-values due to the different parameter values for the interaction effect on the cluster-level that held the effect size of the cluster-level interaction effect constant while modifying the variance of Z_b. The bias was more pronounced (in either direction) with smaller cluster sizes. Once again, when the between-component Z_b did not influence the allocation to treatment groups (in randomized designs), all methods yielded unbiased results. Taking all influential design factors into account, the methods that used the empirical cluster-means and the cluster-mean centered values of the unit-covariate Z as predictors only yielded consistently unbiased results when $ICC(Z) = 0.3$ and average cluster sizes larger or equal than 25. The average biases of the estimator of the average causal effect for all cells of the simulation design are given in Figure 5.3 using the results of the `nlme`-estimator as example. The other methods yielded identical patterns.

Methods that Correct for Unreliability of Z_b. In this section, we will present the bias of the *ACE*-estimator of the two methods that explicitly correct for the unreliability of the empirical cluster means: the adjustment procedure of Croon and van Veldhoven

(2007) and the multigroup multilevel latent variable model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007).

We will begin with the results of the adjustment procedure of Croon and van Veldhoven (2007). The mean bias of the *ACE* averaged over all conditions of the simulation design was 0.015. The average mean bias of the *ACE*-estimator did not vary between the different methods of computing the *ACE* (with the general linear model or via `lace`). The average effect estimators of the two methods were highly correlated over the simulation cells ($r = 0.931$) and the pattern of results was similar. The following detailed description of the results applies to both methods. When $ICC(Z) = 0.3$, the *ACE*-estimator obtained with the adjustment method of Croon and van Veldhoven (2007) was unbiased in all conditions. With smaller $ICC(Z)$-values, a distinctive bias pattern emerged: There was an interaction between the effect size of the cluster-level interaction and the dependency between Z_b and X: Under large effect sizes of the interaction in combination with strong dependencies between the covariate and the treatment variable, a positive bias of the *ACE*-estimator emerged. This bias got more pronounced, the smaller the $ICC(Z)$ got. It was also magnified with smaller average cluster sizes and smaller number of clusters. Once again, across all conditions there was no bias in randomized designs, i.e., if Z_b did not influence the treatment assignment. The mean biases of the *ACE*-estimators obtained with Croon and van Veldhoven's procedure are given in Figure 5.4 using the results obtained with the GLH as example.

In the presentation of the results from the multigroup multilevel latent variable model in `Mplus`, all cells with a sample size of 20 clusters are excluded because of the convergence problems in these conditions. However, the overall pattern of results was not changed when these conditions were included. The average mean bias of the remaining conditions of the experimental design was $MB = 0.001$. Of the remaining 1152 cells, only 124 or 10.76% exhibited a absolute mean bias larger than 0.025. Of these cells, 64 exhibited a positive bias ($MB > 0.025$); 60 showed a negative bias ($MB < -0.025$). The absolute mean bias of the *ACE*-estimator was below the critical value of 0.025 in all conditions when the $ICC(Z) \geq 0.2$. When $ICC(Z) = 0.1$, there were 66 cells with an absolute value of the $MB > 0.025$, when $ICC(Z) = 0.05$, there were 58 cells with an absolute $MB > 0.025$. Closer inspection of these biased cells revealed that small average cluster sizes were particularly critical: The mean bias of the *ACE* estimator became more pronounced when small cluster sizes coincided with strong dependencies between Z_b and X and large treatment-covariate interactions on the cluster-level. This

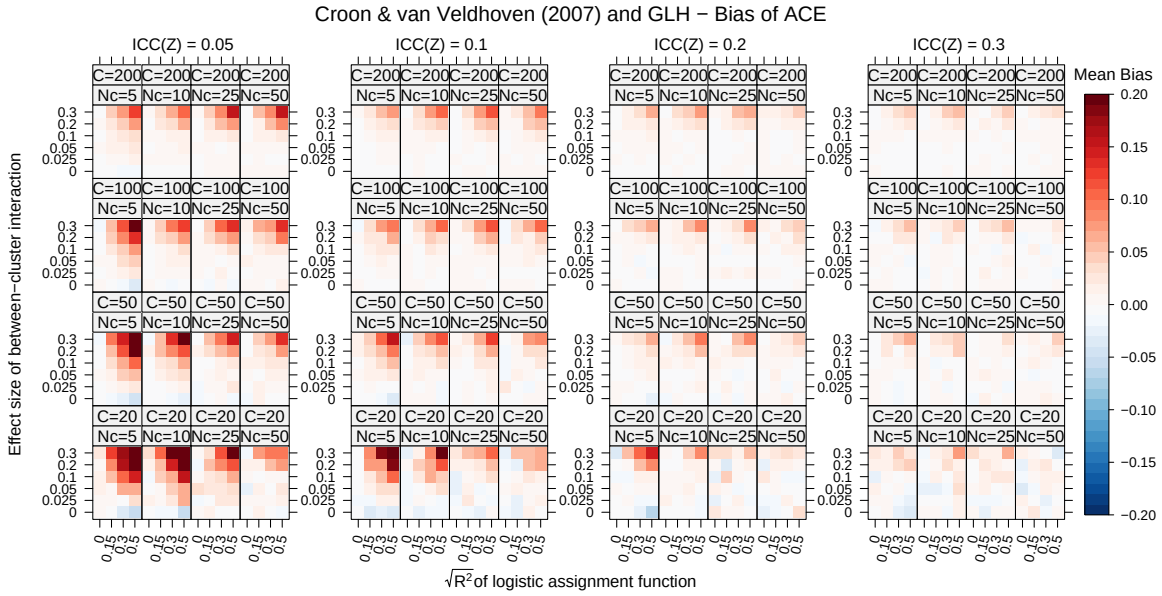

Figure 5.4: Mean bias of *ACE*-estimator: Simple adjustment model implemented with the two-step adjustment procedure of Croon & van Veldhoven (2007)

bias pattern was more pronounced when the $ICC(Z) = 0.05$ and with smaller total number of clusters. The mean biases for all conditions of the simulation design are shown in Figure 5.5 that also includes the conditions where the total number of clusters was 20 that suffered from convergence problems.

Summary. The results of the mean bias of the *ACE*-estimator can be summarized as follows: The naive model implementation in `lace` that modeled only the raw scores of the unit-covariate Z without taking the multilevel decomposition of Z into account led to a strongly biased *ACE*-estimator and is clearly unsuitable in the present context. All methods that modeled the empirical cluster means and within-cluster-residuals resulted in over- and underestimation of the *ACE* consistent with the derivations by Lüdtke et al. (2008). However, the bias was less pronounced than in the naive implementation in `lace`. Since these implementations proved to be relatively robust in terms of convergence — except for the singlegroup model in `Mplus` — a closer look at their standard errors is warranted. The methods that corrected for unreliability of the empirical cluster means as measures of Z_b — the adjustment procedure of Croon and van Veldhoven (2007) and the multigroup multilevel latent variable model in `Mplus` — yielded unbiased *ACE*-estimators with the exception of some extreme conditions combining small

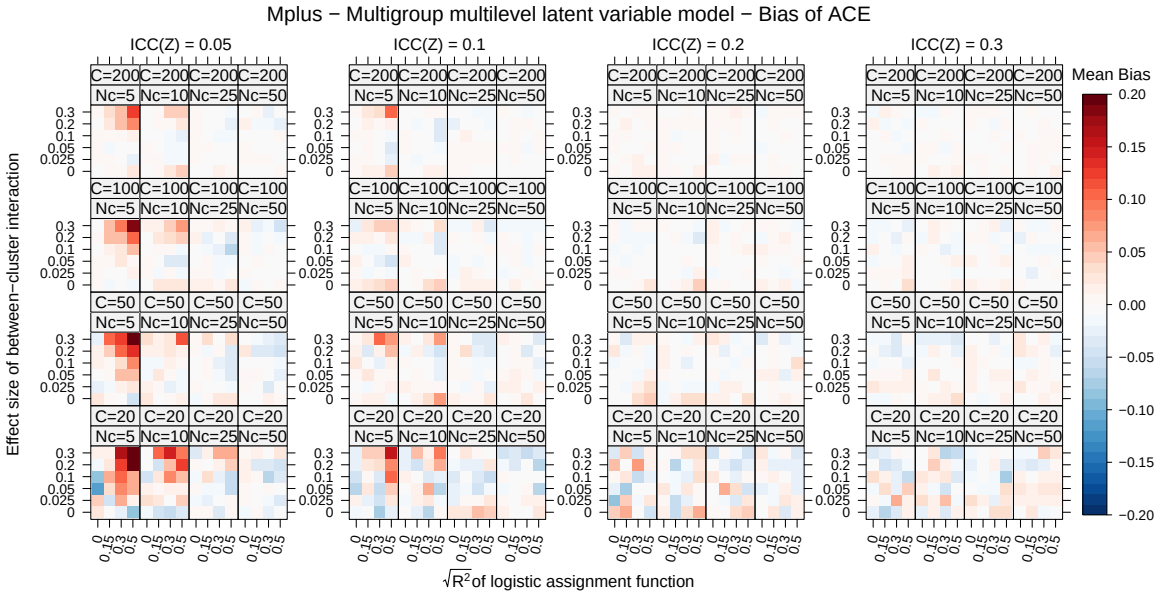

Figure 5.5: Mean bias of *ACE*-estimator: Full adjustment model implemented as multigroup multilevel latent variable model in `Mplus`

clusters, small number of clusters, small $ICC(Z)$-values, strong dependencies of Z_b and X and large effect sizes of the cluster-level interaction.

Relative Bias of Standard Error

In this section, we present the results of the relative bias of the standard error of the *ACE*-estimators of the different implementations of the adjustment model. We start our discussion with the methods that yielded a biased *ACE*-estimator, the simple adjustment model in `nlme`, the full adjustment model in `lace`, `nlme` and as a singlegroup multilevel model in `Mplus`. Then, we describe the results for the methods that yielded unbiased estimators of the average causal effect, the adjustment procedure of Croon and van Veldhoven (2007) and the multigroup multilevel latent variable model in `Mplus`. We will use the mean relative bias (*MRB*) as defined in Equation (D.3) to evaluate the standard error estimators. Following the recommendations by Boomsma and Hoogland (2001), the standard errors were judged as unbiased, if they did not over- or underestimated the variability of the *ACE*-estimates by more than 5% (corresponding to a *MRB* between −0.05 and 0.05).

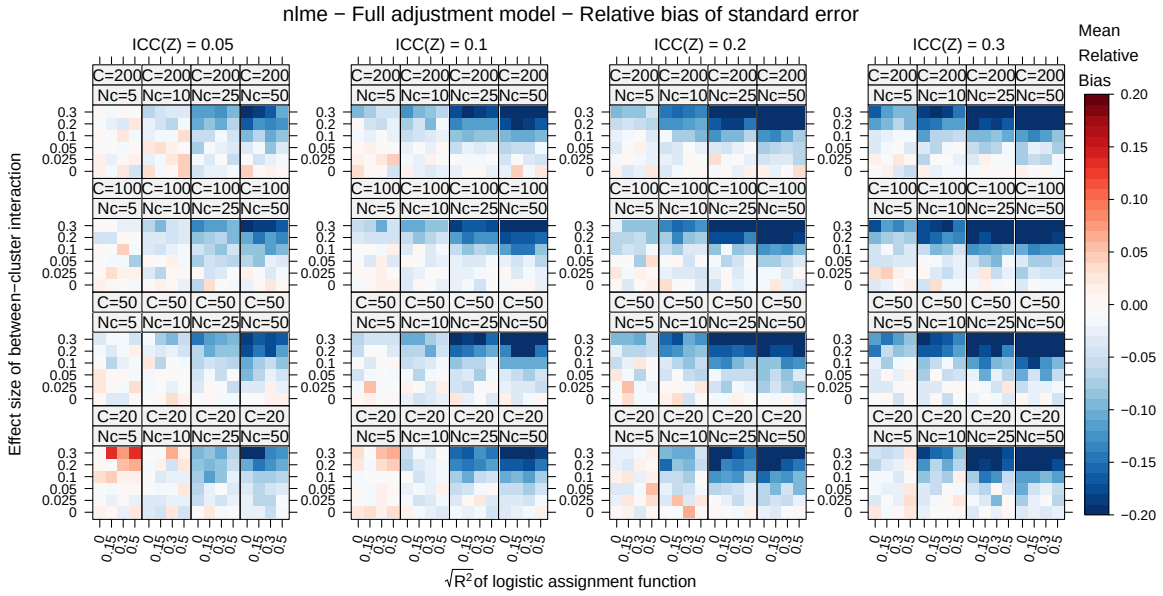

Figure 5.6: Mean relative bias of standard error estimator: Full adjustment model implemented in `nlme`

Methods with Biased *ACE*-Estimators. We start our presentation with the full adjustment model implemented in `lace`. The average *MRB* over all cells of the simulation design was −0.489, indicating that the variability of the *ACE* estimator was underestimated on average by almost 50%. The bias exceeded that cut-off value of 0.05 in all cells.

The *MRB*s for the standard error of the *ACE*-estimator obtained from the simple and the full adjustment model `nlme` were almost identical ($r = 0.996$) and are thus presented concurrently. The average *MRB* of the standard error of the *ACE*-estimator for the two models in `nlme` was equal to −0.062, indicating, on average, an underestimation of the variability of the *ACE* above the absolute cut-off value of 0.05. A total of 657 cells (or 42.77% of all cells) had a *MRB* below the cut-off value $MRB < -0.05$. The *MRB* was affected by the effect size of the cluster-level interaction, the $ICC(Z)$ and the average cluster size. Larger effect sizes of the cluster-level interaction led to a stronger negative bias. This effect was more pronounced, the larger the $ICC(Z)$ and the larger the average cluster size became. There were only 13 cells with an $MRB > 0.05$ mostly in extreme conditions for small total sample sizes. An overview over all cells of the simulation design is given in Figure 5.6 based on the results of the full adjustment model.

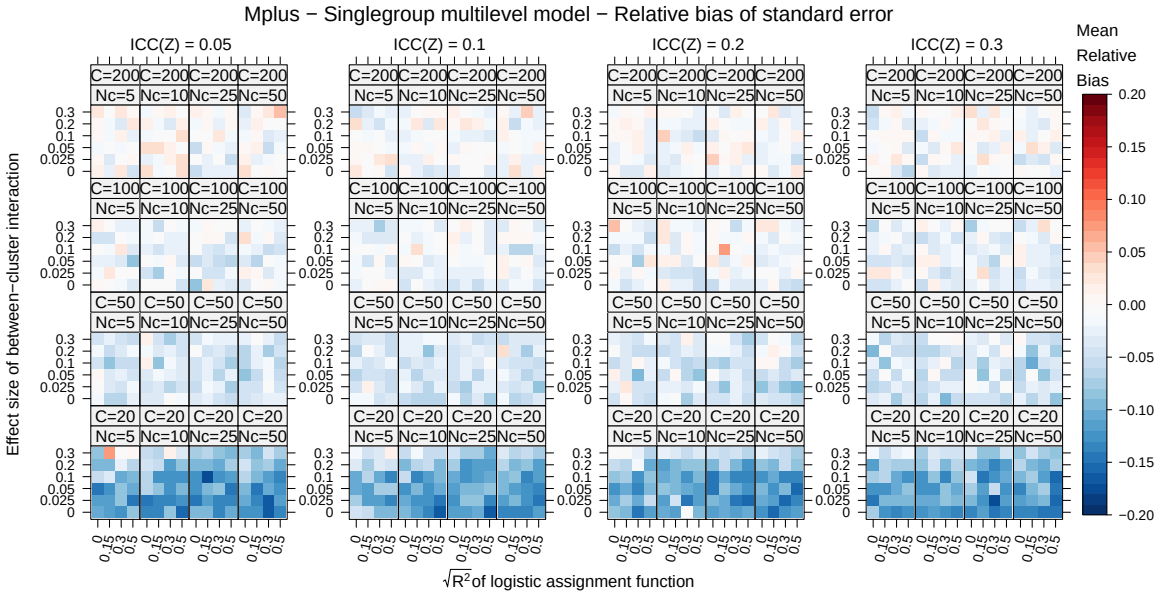

Figure 5.7: Mean relative bias of standard error estimator: Full adjustment model implemented as singlegroup multilevel model in `Mplus`

Due to the convergence problems at the smallest sample of clusters, results for the `Mplus` singlegroup multilevel model are reported without these conditions. In the remaining cells, the average *MRB* was equal to −0.022, indicating a small underestimation of the variability of the corresponding *ACE*. 159 cells or 13.8% of all cells had an *MRB* below the cut-off of $MRB < -0.05$. The average *MRB* was smaller for conditions with 50 clusters ($MRB = -0.038$), than for conditions with 100 cluster ($MRB = -0.019$) or conditions with 200 clusters ($MRB = -0.008$), indicating that the bias disappeared with a larger number of clusters. These results held consistently over all other conditions of the simulation design. Figure 5.7 gives the mean relative biases of the standard error estimator for all cells of the simulation design, including the conditions where the number of clusters was equal to 20 that are not trustworthy due to the convergence problems.

Methods with Unbiased *ACE*-Estimators. We start our discussion with the adjustment method of Croon and van Veldhoven (2007). Standard errors were computed with means of the general linear hypothesis and alternatively using `lace`.

The average mean relative bias of the standard errors computed with the general linear hypothesis was −0.099. The *MRB* was below the cut-off value of −0.05 in a total

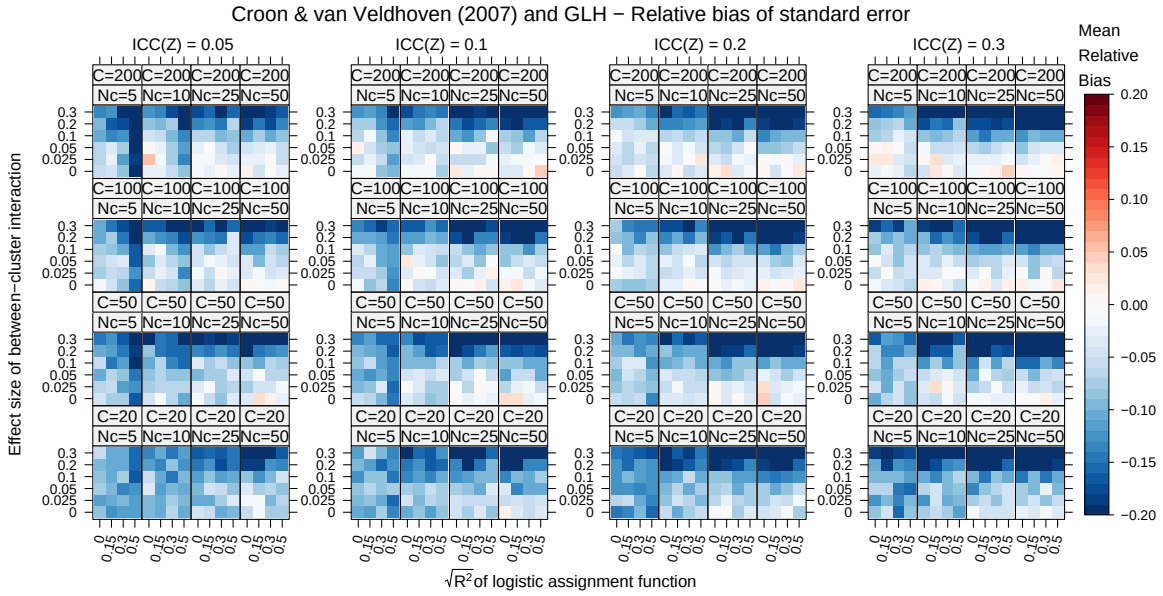

Figure 5.8: Mean relative bias of standard error estimator: Simple adjustment model implemented with the two-step adjustment procedure of Croon & van Veldhoven (2007) using the general linear hypothesis

of 1111 cells of the simulation design, or 72.33% of all conditions, indicating a significant underestimation of the variability of the *ACE*-estimator. The bias was especially pronounced when large interactions of Z_b and the treatment variable X were present. It became also more prevalent under larger values of $ICC(Z)$ and larger cluster sizes. However, a strong negative bias was present throughout all experimental conditions. An overview of the mean relative biases for all cells of the simulation design is given in Figure 5.8.

The average *MRB* of the standard errors for Croon and van Veldhoven's (2007) method obtained with `lace` was −0.067. In 766 cells (49.87% of all cells), the *MRB* was below the threshold of −0.05. Underestimation of the variability of the *ACE*-estimates was especially present for the smallest number of clusters (C=20) and the smallest average cluster size ($\overline{n}_C = 5$). These two effects were further amplified by small values of $ICC(Z)$. In the conditions with larger $ICC(Z)$ values, the bias was attenuated and vanished completely for larger cluster sizes and larger number of clusters. The influence of the effect size of the cluster-level interaction and the dependency between Z_b and X was only discernible for $ICC(Z) = 0.05$ and $ICC(Z) = 0.1$ in connection with small average cluster sizes ($\overline{n}_C = 5$ and $\overline{n}_C = 10$). The mean relative biases for every cell of

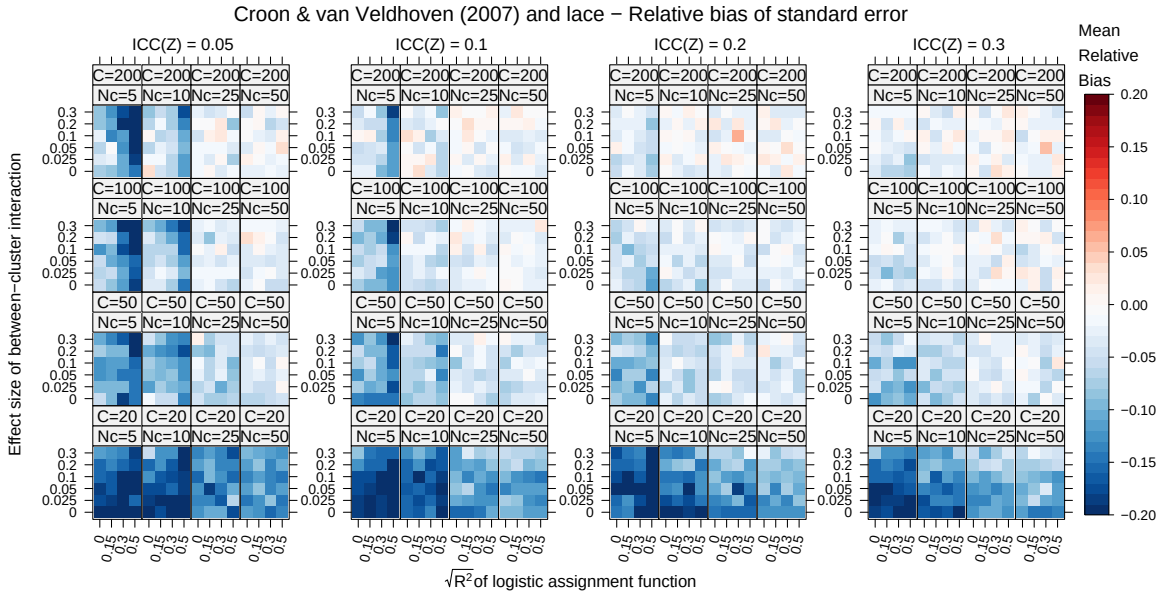

Figure 5.9: Mean relative bias of standard error estimator: Simple adjustment model implemented with the two-step adjustment procedure of Croon & van Veldhoven (2007) using `lace`

the simulation are shown in Figure 5.9.

For the multigroup multilevel latent variable model in `Mplus`, the results for conditions with 20 clusters were discarded from further analysis because of the high-rate of non-converged solutions in these conditions. In the remaining conditions, the average *MRB* of the standard error of the *ACE*-estimator was 0.042, indicating an overestimation of the variability of the *ACE*-estimator. A total of 468 cells (36.63% of the remaining cells) had an absolute *MRB* above the cut-off value of 0.05. The standard error overestimated the variability of the *ACE*-estimator in 422 cells, and underestimated it in only 46 cells. Further inspection of the results indicated that the overestimation was especially pronounced for the smallest average cluster sizes and further elevated under small *ICC*(*Z*)-values. Larger values of the *ICC*(*Z*) reduced the bias across all other conditions, although in conditions that combined small average cluster sizes with a small number of clusters, overestimation of the variability of the *ACE*-estimator was still present. All mean relative biases of the estimator of the standard error of the average causal effect are shown in Figure 5.10 including the conditions with 20 clusters that were hampered by convergence problems.

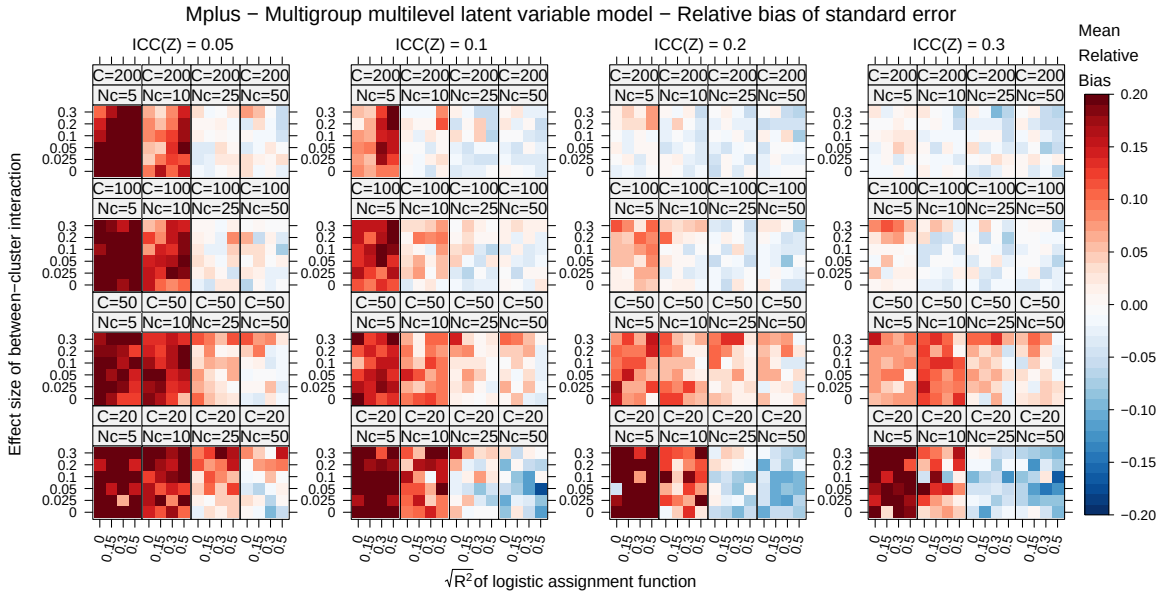

Figure 5.10: Mean relative bias of standard error estimator: Full adjustment model implemented as multigroup multilevel latent variable model in `Mplus`

Summary. Among the methods with biased estimation of the *ACE*, the `lace`-implementation of the full adjustment model clearly performed worst and yielded standard errors that markedly and expectedly underestimated the variability of the corresponding *ACE*-estimator. The multilevel implementation of the adjustment model in `nlme` also underestimated the variability of its *ACE*-estimator when strong interactions between X and Z_b were present. Among these methods the `Mplus` singlegroup multilevel model that treats the covariate Z as a stochastic predictor gave the most accurate standard error estimator – at least after excluding all conditions with 20 clusters, where this model had convergence problems.

The implementations of the adjustment model that resulted in unbiased *ACE*-estimators performed differently with respect to the accuracy of the standard error as a measure for the variability of the *ACE*-estimator. The results for the adjustment procedure of Croon and van Veldhoven (2007) differed depending on the implementation of the test: When the general linear hypothesis was used to compute the standard error, the variability of the *ACE* was underestimated especially with small number of clusters. When `lace` was used to obtain the standard errors, the underestimation was less pronounced and but still present under small $ICC(Z)$, a small number of clusters and

small average cluster sizes. The standard error estimator of the multigroup multilevel latent variable model in `Mplus` overestimated the variability of the corresponding *ACE*-estimator in some conditions: This bias was especially pronounced under small $ICC(Z)$ and for small average cluster sizes.

Type-1-Error Rate

In the following section, we will present the results for the empirical type-1-error rates for two-sided-significance tests of the H_0: $ACE_{10} = 0$ at a significance level of 0.05. Tests at other significance levels yielded similar results. The presentation of the results will be structured similarly to the presentation of the mean relative biases of the standard error estimators. We will start with discussing the methods with a biased *ACE*-estimator, the adjustment model as implemented in `lace`, in `nlme` and as singlegroup multilevel model with manifest variables in `Mplus`. We will then present the results of the methods with an unbiased *ACE*-estimator, the adjustment procedure of Croon and van Veldhoven (2007) and the multigroup multilevel latent variable model in `Mplus`. In line with the suggestions by Boomsma and Hoogland (2001), the limits of the 95%-confidence interval for the rejection frequency of an adequate significance test were calculated (for details see Appendix D) and used to evaluate the performance of the significance tests. The lower limit of the confidence interval for 1000 replications was 0.037; the upper limit was 0.064.

Methods with Biased *ACE*-Estimator. We will begin with presenting the results of the implementation of the full adjustment model in `lace`. The mean empirical type-1-error rate over all conditions was 0.410 and thus clearly too liberal. Further inspection of the results indicated that even the smallest empirical type-1-error rate was equal to 0.106 and thus exceeded the nominal α-level by far, rendering a detailed analysis unnecessary.

Once more, the implementations of the simple and the full adjustment model in `nlme` gave almost identical results: They yielded an average empirical type-1-error rate of 0.089 and thus overall a slightly too liberal test of the null hypothesis of no average causal effect. Further inspection of the results revealed that a total of 1039 cells (67.64% of all cells) had type-1-error rates outside of the 95%-confidence interval — indicating too liberal significance tests. The exceedance was especially pronounced in

the presence of large interactions on the cluster-level (and for small values of $ICC(Z)$ also in combination with large $\sqrt{R^2}$ of the logistic assignment function) – exactly in the conditions that had either resulted in biased estimators of the average causal effect or an underestimation of the variability of the average causal effect by the standard error estimator.

In the analysis of the results of the `Mplus` singlegroup multilevel model with cluster means and group-mean-centered scores of the covariate Z as predictors, the conditions with 20 clusters that exhibited significant convergence problems were omitted. The average type-1-error rate in the remaining conditions was 0.074. Closer inspection of the results revealed that 498 cells (or 43.22%) yielded a type-1-error rate outside of the 95%-confidence interval. The significance test proved to be too liberal in the conditions that had exhibited a strong bias of the average causal effect: Under large interactions, small values of $ICC(Z)$ and large $\sqrt{R^2}$ of the logistic assignment function, the empirical type-1-error rates exceeded the nominal significance level most strongly. Independent of these combined effects, the significance test proved to be too liberal for 50 clusters, the smallest remaining number of clusters.

Methods with Unbiased *ACE*-Estimator. The adjustment procedure of Croon and van Veldhoven (2007) in combination with the general linear model yielded an average type-1-error rate of 0.075. 775 cells (50.46%) exceeded that upper limit of the 95%-confidence interval, indicating that this procedure yielded too liberal significance tests of the null hypothesis of no average causal effect. The significance test was especially liberal in conditions with large interactions and larger average cluster sizes – the conditions in which the standard error was negatively biased. When the $ICC(Z)$ was equal to 0.05, the conditions with a large dependency between Z_b and the treatment variable X similarly exhibited heightened type-1-error rates – these conditions had yielded the strongest positive bias of the average causal effect estimator.

When the adjustment procedure of Croon and van Veldhoven (2007) was used in combination with `lace` to test the null hypothesis of no average causal effect, the average type-1-error rate was 0.072. 803 cells (52.28%) exceeded the upper limit of the 95%-confidence interval. When only conditions with more than 20 clusters were considered this number dropped to 420 (36.46%), but still indicated too liberal tests of the null hypothesis. Heightened type-1-error rates were especially prevalent for small number of clusters, where standard error estimators underestimated the variability of

the *ACE*-estimator, large interactions and strong dependencies between X and Z_b. The latter two effects were due to the bias of the *ACE*-estimator in these conditions and vanished with larger values of $ICC(Z)$.

In the analysis of the multigroup multilevel latent variable model in `Mplus`, all conditions with 20 clusters were dropped from the analysis due the convergence problems in these cells. The analysis of the remaining conditions yielded an average type-1-error rate of 0.052 at the nominal significance level of $\alpha = 0.05$. Further inspection of the results indicated that only 742 cells (equal to 64.41% of all remaining conditions) were within the 95%-confidence interval of the type-1-error rate. In 172 cells (corresponding to 14.93% of all cells) the observed type-1-error rates were below the lower limit of the confidence interval ($\alpha < 0.037$) with the minimum observed type-1-error rate at 0.013, indicating a conservative test of the null hypothesis. 229 cells (19.88%) yielded a type-1-error rate that exceeded that upper limit of the confidence interval ($\alpha > 0.064$) with a maximal observed type-1-error rate at 0.105, indicating too liberal significance tests in these cells. In line with the mean relative bias of the standard errors, conservative performance of test was prevalent in cells with a small $ICC(Z)$ in combination with small average cluster sizes. Liberal tests were found for 50 clusters and large values of $ICC(Z)$.

Summary. Unsurprisingly, the empirical type-1-error rates closely mirrored the bias of the *ACE*-estimator and the results of the mean relative bias of the corresponding standard error estimators: Two-sided significance tests of the null hypothesis of no average causal effect ($H_0 : ACE = 0$) at an α-level of 0.05 were too liberal when the standard error estimator underestimated the empirical variability of the *ACE*-estimator. The significance tests proved to be too conservative when the standard error estimator overestimated the empirical variability of the corresponding *ACE*-estimator. A biased *ACE*-estimator also led to higher rejection rates of the null hypothesis and to significance tests that were too liberal.

The most liberal tests were obtained with the implementation of the adjustment model in `lace`. This implementation did not account for the multilevel structure of the data and proved once more to be unsuitable for the analysis of average causal effects in multilevel designs. The adjustment model in `nlme` led to liberal tests in the conditions where the standard errors were underestimated. This effect was magnified in the conditions, in which the *ACE*-estimator exhibited a negative bias. The singlegroup

multilevel implementation in `Mplus` also led to slightly progressive significance tests, once again the effect was pronounced in the conditions with a biased *ACE*-estimator.

The multigroup multilevel mode in `Mplus` yielded conservative significance tests in those conditions where the standard error underestimated the variability of the *ACE*-estimator, notably for small values of $ICC(Z)$ in combination with small cluster sizes. At larger $ICC(Z)$-values, the nominal significance levels were kept or were slightly too liberal. The adjustment method of Croon and van Veldhoven (2007) in combination with the general linear model led to liberal significance tests in the conditions where the standard errors had a negative relative bias. In combination with `lace`, the adjustment procedure yielded accurate significance tests except for some combinations of small $ICC(Z)$, large interactions, strong dependencies between the cluster-component Z_b and the treatment variable X and small average cluster sizes.

Efficiency of *ACE*-Estimator

Finally, we compared the mean-squared errors (*MSE*) of the *ACE*-estimators focusing on two questions: (1) The comparison of the efficiency of the simple and full adjustment model in `nlme` and (2) the comparison of the efficiency of the multigroup multilevel latent variable model in `Mplus`, the hierarchical linear model in `nlme`, the singlegroup multilevel model in `Mplus` and the adjustment procedure of Croon and van Veldhoven (2007).

The efficiencies of the simple and the full adjustment model in `nlme` were compared by dividing the *MSE* of the full model by the *MSE* of the simple model. The two *MSE*s were highly correlated ($r = 0.999$). The average *MSE*-ratio was 0.999 — indicating that the two methods were on average equally efficient in estimating the *ACE*. In 865 cells (43.68%), the *MSE* ratio was below one; in the remaining 671 cells (56.32%), the *MSE* ratio was above one.

The multigroup multilevel latent variable model in `Mplus` was used as the reference method to compare the efficiency of the other statistical models, because it proved to be the most promising method combining an unbiased *ACE*-estimator and an unbiased (or at least not negatively biased) standard error estimator. All comparisons for the `Mplus` models omitted conditions with 20 clusters, because of the convergence problems present in these conditions. The *MSE* of the `Mplus` singlegroup multilevel model and the hierarchical linear model in `nlme` were highly correlated ($r = 0.999$) and the

average ratio was close to 1 ($\bar{x} = 0.999$) with very small variation ($\hat{\sigma} = 0.005$). Therefore, the presentation of the results will be confined to the `nlme` estimator in the remainder, since comparisons with the `Mplus` singlegroup multilevel estimator yielded almost identical results.

The comparison of the *MSE* of the *ACE*-estimator obtained with the `Mplus` multigroup multilevel latent variable model and *ACE*-estimator obtained with `nlme` was done by dividing the *MSE* of the `nlme` model by the *MSE* of the `Mplus` model. The mean *MSE* ratio was equal to 0.890, indicating that on average the `nlme` estimator was more efficient. A closer inspection of the results indicated that this advantage was especially pronounced for the conditions with small $ICC(Z)$ and in conditions with small average cluster sizes and small number of clusters. The differences between small and large average cluster sizes were greatly reduced in conditions with large $ICC(Z)$-values and both methods were roughly equivalent in their efficiency. There were some conditions in which the `Mplus` estimator was clearly more efficient than the `nlme` estimator: These were the cells in which the bias of the `nlme` estimator was strong, notably in some conditions with strong dependencies and large interactions.

The *MSE* of the multigroup multilevel model in `Mplus` and the adjustment procedure of Croon and van Veldhoven (2007) revealed a large advantage in efficiency for the latter method: The mean *MSE*-ratio was 0.486 and all simulation cells resulted in *MSE*-ratios smaller than 1. Identical results were obtained no matter which method was used to obtain the *ACE*-estimator.

Summary of Results

The results of the simulation can be summarized with respect to the research questions introduced in Section 5.2.2 as follows:

1. Neglecting the multilevel decomposition of the unit-covariate Z and the differential effects of the between-component Z_b and the within-component Z_w on the outcome variable Y and including only the unit-covariate Z in an adjustment model was clearly not an appropriate method to obtain an *ACE*-estimator: The naive model implementation in `lace` yielded a biased *ACE*-estimator in almost all simulation conditions.

2. Even if the multilevel decomposition of the unit-covariate Z was appropriately

accounted for in a singlelevel adjustment model in `lace` that modeled both the within- and the between-component of the unit-covariate Z with the empirical cluster-means and cluster-mean centered scores of Z, the resulting standard error showed a considerable bias. As expected, the omission of the variance components $Var(r_{0;C})$ and $Var(r_{10;C})$ for residuals at the cluster-level from the model led to a severe underestimation of the empirical variability of the *ACE*-estimator and to too liberal significance tests.

3. All statistical models that did not account for the fact that cluster means were fallible measures of the underlying values of the regression Z_b yielded biased *ACE*-estimators. Both the multigroup multilevel latent variable model in `Mplus` (Lüdtke et al., 2008) and the adjustment procedure of Croon and van Veldhoven (2007) corrected this bias. While the former yielded correct standard errors (albeit with an overestimation in for small numbers of clusters and small $ICC(Z)$-values), it was hampered by convergence problems for the smallest number of clusters. The adjustment procedure of Croon and van Veldhoven (2007) did not have any convergence problems, but yielded standard errors that underestimated the variability of the estimator even if the *ACE*-estimator and its standard error were obtained with `lace`.

4. The *ACE*-estimators obtained from the simple and the full adjustment model within the same statistical framework were almost identical. Both models shared the problems — biased *ACE*-estimator and standard errors respectively — of the corresponding framework. In the context of the simulation study, no clear efficiency advantage emerged for either model: The implementations of the two models in `nlme` yielded almost identical results. However, both the simple adjustment model as implemented in the adjustment procedure of Croon and van Veldhoven (2007) and the full adjustment model as implemented in the multigroup multilevel latent variable model in `Mplus` gave unbiased *ACE*-estimators.

5. In line with previous results for singlelevel models (Kröhne, 2009; Nagengast, 2006), we also found that standard errors were underestimated when the unit-covariate Z was not treated as a stochastic predictor — even if the other variance components $Var(r_{0;C})$ and $Var(r_{10;C})$ were modeled appropriately. Only the implementations of the adjustment model in the multilevel structural equation model

in `Mplus` yielded unbiased standard errors. The conventional hierarchical linear model implementation in `nlme` that treats all predictors as fixed yielded a negatively biased standard error.

In the next section, we will apply the different implementations of the generalized ANCOVA to an illustrative example. We will further discuss the results of the simulation study in Section 5.4.

5.3 Example Analysis

In this section, we illustrate the statistical implementations of the generalized ANCOVA for designs with treatment assignment at the cluster-level with an empirical example from the first-wave of the Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K, National Center for Education Statistics, 2001). We will estimate the average effect of the presence of a mud-play area in kindergarten classrooms on quantitative skills at the end of the kindergarten year controlling for the between- and within-effect of the pre-test in quantitative skills taken at the beginning of the Kindergarten year. Again, the analyses are only intended to demonstrate the importance of adjusting average treatment effects for the influences of covariates in multilevel observational studies and quasi-experiments, accounting for the decomposition of unit-covariates into within- and between-components and the flexibility of the generalized ANCOVA in doing so; they are not aimed at deriving substantive insights into effects of an enriched classroom environment on quantitative skills during kindergarten. Conceptually, the data comes from a quasi-experiment with treatment assignment at the cluster-level, using pre-existing clusters and self-selection to treatment conditions. The interpretation of the obtained effects as average causal effects rests on the assumption of an unbiased cluster-covariate-treatment regression $E(Y \mid X, Z_b)$ — an assumption that is not tested explicitly in the analyses.

5.3.1 Methods

Design

The Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K) was a multisource, multimethod study in the US that focused on children's early school

experiences beginning with kindergarten. The following description of its design and the employed methods are taken from the userguide for the base year public use data files (National Center for Education Statistics, 2001). ECLS-K followed a nationally representative cohort of 21260 children from kindergarten to fifth grade. Here, we are using data from the base year, collected in the fall of 1998 and the spring of 1999, and study the effects of a kindergarten-level treatment on kindergarten outcome. We are using a sample of 17151 children from 2930 classrooms with complete data for all considered variables.

ECLS-K used a complex sample design to obtain a nationally representative sample of children (see National Center for Education Statistics, 2001). Analyzing the data with the goal of obtaining nationally representative results would require a complex weighting scheme to account for design effects, non-response and dropout. Since we only use this data to illustrate different implementations of the adjustment model, we ignore this additional complexity and analyze the data as if it was obtained from a simple random sample.

Materials

The assessment of quantitative skills took place at the beginning and near the end of the kindergarten year using computer-assisted personal interviewing (CAPI) and the same assessment material at both timepoints. Assessment sessions lasted 50 to 70 minutes and included a variety of other cognitive and non-cognitive assessments. The items of the quantitative skills assessment were designed to measure skills in conceptual knowledge, procedural knowledge and problem solving. A two-stage assessment design was implemented: Children first completed a 12 item routing test that was followed by one of several alternative second-stage forms that varied in difficulty and were chosen according to the performance in the routing test. Individual scores for the pre- and the posttest on a common scale were obtained using IRT procedures (National Center for Education Statistics, 2001). The average reliability of the IRT-scores — computed as 1 minus the ratio of the average measurement error variance to the total variance — was 0.92 in the fall assessment and 0.94 in the spring assessment. The descriptive statistics for the pre- and posttest measures of quantitative skills in our sample are given in Table 5.4.

Kindergarten teachers completed a self-administered questionnaire including ques-

tions about classroom characteristics, educational activities in their classroom, school environment and school climate as well as questions about the sampled children in their classroom. The teacher identification variable at the fall assessment was used as cluster variable in the analysis. There were 2930 classrooms in the analysis. The average classroom size was 5.854. The smallest classroom contained only 1 student, the largest classroom contained 27 students. The empirical cluster means of the pre-test score in mathematics knowledge were computed by averaging over the pre-test values of the students associated with a teacher. Since not all students of a classroom were sampled, the empirical cluster means were fallible measures of the latent classroom variable Z_b (cf., Lüdtke et al., 2008). Each teacher indicated whether or not his classroom was equipped with a mud-play area. This variable was used as treatment variable on the cluster-level whose average causal effect on mathematics knowledge was studied controlling for the effects of the between-component Z_b, the within-component Z_w and their product. 1427 classrooms were equipped with a mud-play area (corresponding to 8283 students), the remaining 1503 classrooms (corresponding to 8868 students) were not equipped with a mud-play area, resulting in roughly equal sample sizes of both the treatment and the control group.

Statistical Procedures

We compared the following implementations of the generalized ANCOVA for designs with treatment assignment at the cluster-level. The models are the same as in the simulation study. Their full description is given in Appendix B. Specifically, we estimated the *ACE* with

(a) the naive model in `lace`,

(b) the full adjustment model in `lace`,

(c) the simple adjustment model in `nlme`,

(d) the full adjustment model in `nlme`,

(e) the full adjustment model using the multigroup multilevel latent variable model in `Mplus`,

(f) the simple adjustment model using Croon and van Veldhoven's (2007) procedure.

Table 5.4: Descriptive statistics of the variables in the ECLS-K data set

Variable	Mean	*SD*	*ICC*	Min	Max
Math Pre-Test	19.539	7.353	0.242	6.651	59.815
Math Post-Test	27.675	8.864	0.258	7.537	59.339
Treatment	0.483	0.499	1	0	1

Additionally and for comparison, we also computed the unadjusted treatment effect in `nlme`. As suggested by Steyer and Partchev (2008), effect sizes $d(\widehat{ACE})$ for the average effect estimates were obtained by dividing the estimate with the standard deviation of the outcome variable in the control group. For the simple adjustment model of Croon and van Veldhoven (2007), the standard deviation of the adjusted cluster means of the outcome variable were used for this purpose. Estimates of the intraclass correlation coefficients (*ICC*) of the pre- and the post-test were obtained from intercept-only models for the pre- and post-test measures of mathematical achievement specified in `nlme`. The estimate of the dependency of the treatment variable X and the cluster-means of the between-component Z_b on the treatment probabilities were obtained with a logistic mixed regression with the `R`-package `MASS` (Venables & Ripley, 2002).

5.3.2 Results

Intercept-Only Models. The intercept-only model for the pre-test measure of quantitative skills showed that a significant amount of variance of the pre-test was located between classrooms ($\hat{\sigma}^2_{Z_b} = 13.084; \hat{\sigma}^2_{Z_w} = 41.070; \widehat{ICC}(Z) = 0.242$). The estimated intercept parameter was $\hat{\gamma}_{00} = 19.101$ ($SE = 0.088, t = 217.057, p < 0.001$). The average reliability of the cluster means as measures for the true values of the between-component Z_b was 0.651.

The intercept-only model for the post-test measure of quantitative skills also showed that a significant amount of variance of the post-test was located between classrooms ($\hat{\sigma}^2_{Z_b} = 20.339; \hat{\sigma}^2_{Z_w} = 58.579; \widehat{ICC}(Z) = 0.258$). The estimated intercept parameter was $\hat{\gamma}_{00} = 27.2508$ ($SE = 0.107, t = 254.680, p < 0.001$), roughly eight points higher than the pre-test value.

Dependency of Covariate and Treatment. The logistic regression of the treatment variable on the cluster-means of the mathematics pre-test indicated that the between-component of the pre-test significantly influenced treatment assignment probabilities ($\beta = 0.032, SE = 0.008, Z = 4.320, p < 0.001$). The value of Nagelkerke's (1991) $\sqrt{R^2}$ was 0.093.

Unadjusted Treatment Effect. The unadjusted treatment effect obtained with `nlme` was $\hat{\gamma}_{10} = 0.540$ ($SE = 0.215, t = 2.509, p = 0.012$), indicating a small positive effect [$d(\widehat{ACE}) = 0.061$] of the presence of a mud-play area in the kindergarten classroom on quantitative skills, if no other covariates were considered. This supposed effect was statistically significant at a two-tailed significance-level of 0.05. For comparison purposes, the treatment effects of all implementations are displayed in Table 5.5 on the following page.

Naive Adjustment Model in `lace`. The naive adjustment model in `lace` that only controlled for the raw scores of the pre-test and did not take the multilevel structure of the data into account estimated the average causal effect of the treatment as $\widehat{ACE} = -0.124$ ($SE = 0.077, t = -1.602, p = 0.110$). This result indicated a small negative average effect of the treatment ($d(\widehat{ACE}) = -0.014$) that was not statistically significant at a two-tailed significance-level of 0.05.

Full Adjustment Model in `lace`. The full adjustment model in `lace` that included the cluster-means of the pre-test, the cluster-mean centered values of the pre-test and their interactions as predictors, but not the additional variance components for the intercept, estimated the average causal effect of the treatment as $\widehat{ACE} = -0.156$ ($SE = 0.077, t = -2.026, p = 0.042$). This result indicated a small negative average effect of the treatment [$d(\widehat{ACE}) = -0.018$] that was statistically significant at a two-tailed significance-level of 0.05.

Table 5.5: Comparison of the estimated *ACE*s of the different adjustment procedures

Method	$\widehat{ACE}$	$d(\widehat{ACE})$	*SE*	*t*-value	*p*-value	95%-conf.-interval
`nlme`: No adjustment	0.540	0.061	0.215	2.510	0.012	[0.188; 0.962]
`lace`: Naive model	−0.124	−0.014	0.077	−1.602	0.110	[−0.276; 0.028]
`lace`: Full model	−0.156	−0.018	0.077	−2.026	0.042	[−0.307, −0.005]
`nlme`: Simple model	−0.156	−0.018	0.114	−1.370	0.171	[-0.379; 0.067]
`nlme`: Full model	−0.160	−0.018	0.104	−1.538	0.124	[−0.363; 0.044]
`Mplus`: Full model	−0.231	−0.026	0.105	−2.190	0.048	[−0.437; −0.025]
Croon + GLH: Simple model	−0.055	−0.018	0.028	−1.964	0.049	[−0.110; −0.0001]
Croon + `lace`: Simple model	−0.054	−0.018	0.028	−1.943	0.051	[−0.109; 0.0004]

Simple Adjustment Model in `nlme`. The simple adjustment model in `nlme` that included only the cluster-means of the pre-test, the treatment variable and their interaction as predictors, estimated the average causal effect of the treatment with the general linear hypothesis as $\widehat{ACE} = -0.156$ ($SE = 0.114, t = -1.370, p = 0.171$). The simple adjustment model in `nlme` indicated a small negative average treatment effect [$d(\widehat{ACE}) = -0.018$] that was not statistically significant at a two-tailed significance-level of 0.05.

Full Adjustment Model in `nlme`. The full adjustment model in `nlme` included the cluster-means of the pre-test, the cluster-mean centered scores of the pre-test, the treatment variable and their interactions as predictors — as in the simulation study, a random effect of the intercept was included.

All parameters of the full adjustment model in `nlme` are given in Table 5.6 for comparison purposes and to further characterize the data. The results indicated that the effects of the within-component Z_w, the between-component Z_b and their interaction in the control group were significant at the 5%-two-sided significance level. The effect of the between-component Z_b remained constant in the treatment group, while the effects of the within-component Z_w and of the interaction between Z_b and Z_w were different in the treatment group as indicated by the significant two- and three-way interactions. The residual intraclass correlation coefficient *rICC* of the outcome variable conditional on the covariates was equal to 0.133.

The average causal effect of the treatment estimated by this model with the general linear hypothesis was $\widehat{ACE} = -0.159$ ($SE = 0.104, z = -1.531, p = 0.124$). Thus, the full adjustment model in `nlme` indicated a small negative average treatment effect [$d(\widehat{ACE}) = -0.018$] that was not statistically significant at a two-tailed significance-level of 0.05.

Full Adjustment Model in `Mplus`. The implementation of the full adjustment model in the multigroup multilevel latent variable model in `Mplus` that decomposed the pre-test into its latent between- and within components and modeled the effects of these covariates and their product separately for each treatment group, estimated the average causal effect of the treatment as non-linear constraint of the model parameters and the expected value of the covariate. The estimated ACE was $\widehat{ACE} = -0.231$ ($SE = 0.105, z = -2.190, p = 0.048$). The full adjustment model in `Mplus` indicated a

Table 5.6: Model parameters of the full adjustment model in `nlme` for the ECLS-K data set

Parameter	*Estimate*	*SE*	*t*-value	95%-conf.-interval
Fixed Effects				
γ_{00}: Intercept	6.742	0.302	22.303	[6.150; 7.334]
γ_{01}: Z_b	1.077	0.016	69.487	[1.046; 1.108]
γ_{04}: Z_w	1.557	0.045	34.969	[1.469; 1.645]
γ_{05}: $Z_w \cdot Z_b$	−0.029	0.002	−14.205	[−0.033; −0.025]
γ_{10}: X	−0.013	0.440	−0.029	[−0.875; 0.849]
γ_{11}: $X \cdot Z_b$	−0.007	0.022	−0.336	[−0.050; 0.036]
γ_{14}: $X \cdot Z_w$	−0.201	0.065	3.100	[−0.328; −0.074]
γ_{15}: $X \cdot Z_w \cdot Z_b$	0.010	0.003	3.436	[0.004, 0.016]
Variance Components				
Level 1				
σ^2_ε	21.638			
Level 2				
$\sigma^2_{u_0}$	3.236			

small negative average causal effect estimator [$d(\widehat{ACE}) = -0.026$] that was statistically significant at a two-tailed significance-level of 0.05.

Simple Adjustment Model using Croon and van Veldhoven's (2007) procedure. The implementation of the simple adjustment model using the modified version of Croon and van Veldhoven's (2007) adjustment procedure that corrected the cluster-means of the unit-covariate for their unreliability estimated the average causal effect of the treatment with the general linear hypothesis as $\widehat{ACE} = -0.055$ ($SE = 0.028, z = -1.964, p = 0.049$). The small negative average effect [$d(\widehat{ACE}) = -0.018$] was statistically significant at a two-tailed significance-level of 0.05. Similar results emerged when `lace` was used to estimate the *ACE* with the adjustment procedure of Croon and van Veldhoven: The estimated average causal effect was equal to $\widehat{ACE} = 0.054$ ($SE = 0.028, z = -1.943, p = 0.051$). However, the effect estimated with `lace` was not statistically significant at a two-tailed significance-level of 0.05.

5.3.3 Discussion

The data set and the results of the example analysis were similar to the results from the simulation study in many ways. The actual parametric conditions of the sample data were not explicitly included in the simulation design, a comparable condition would have been found with 200 clusters, an average cluster size of 5, *ICC* of the unit-covariate between 0.2 and 0.3, the $\sqrt{R^2}$ of the logistic assignment function at 0.15 and no interaction effect at the between-cluster level. As in the simulation study, the regression weight of the within-component Z_w was numerically larger than the regression weight of the between-component Z_b in the control group. The interaction of the within-component Z_w and the treatment variable X had a negative regression weight; the regression weight of the interaction of the between-component Z_b and the treatment variable X was zero. The residual intraclass correlation coefficient *rICC* of the outcome variable was substantial and only slightly smaller than in the simulation study. The three-way interaction between the treatment variable Z_b and Z_w was also a significant predictor of the outcome variable. As in the simulation study, the empirical cluster means could be conceived of as fallible measures of the true average values of the pre-test, since they were not calculated from all students within the respective classroom. However, the total sample size and the number of clusters in the ECLS-K data set were considerably larger than in any condition of the simulation design and the cluster sizes had a larger variance than in the simulation.

Judging from the results of the comparable condition in the simulation study, the *ACE*-estimate obtained with the multigroup multilevel model in `Mplus` would have been expected to be unbiased, as would the corresponding estimate of the standard error. The *ACE*-estimate obtained with the full model in `nlme` would be expected to exhibit a small negative bias, its standard error should adequately reflect the variability of the estimator. Similarly, the *ACE*-estimate obtained from the adjustment procedure of Croon and van Veldhoven (2007) would be expected to be unbiased with an unbiased standard error. In addition, it could be expected that this estimate would be more efficient than the *ACE* obtained from `Mplus`, i.e., to show less variability and be on average closer to the true value of the *ACE*. Both estimators obtained with `lace` were expected to have negatively biased standard errors. The *ACE*-estimator from the naive model should show a small positive bias, while the *ACE*-estimator from the full model was expected to slightly overestimate the true *ACE*.

The empirical results were in line with the predictions from the simulation study: Overall, the results from the appropriate adjustment models indicated that the *ACE* was on the boundary of statistical significance at a significance-level of 0.05. The upper bound of the 95%-confidence limit was close to zero for all estimates. In line with the results on the relative efficiency of the estimators, the standard error of the multigroup multilevel latent variable model in `Mplus` had the largest standard error — albeit only slightly larger than the `nlme` model — and yielded the most negative *ACE*-estimate. The simple adjustment model in `nlme` had a larger standard error, indicating that excluding the within-component Z_w from the model decreased the efficiency of the *ACE*-estimate. The adjustment procedure of Croon and van Veldhoven (2007) had the smallest standard error and yielded the *ACE*-estimate, the closest to zero. If the *ACE*-estimates were converted into the proper effect size metric, the results indicated that all average effects were highly similar — with the only exception of the `Mplus` estimate that was slightly, though not substantially smaller. The `Mplus` estimate should, however, not be carelessly dismissed: The average cluster sizes were relatively small; consequently the resulting reliability of the cluster means was relatively low, implying that a substantial bias was to be expected for the methods that did not correct for this unreliability. Additionally, there was an interaction between the treatment variable X, the between-component Z_b and the within-component Z_w that was not captured by Croon and van Veldhoven's (2007) adjustment procedure and was adequately taken into account only by the `Mplus` multigroup multilevel latent variable model.

5.4 Discussion

In this section, we will discuss the adjustment models for multilevel designs with treatment assignment at the cluster-level, their implementation in statistical models, the results of the simulation study and the example analysis. We first focus on the most promising statistical implementations and link the results of the simulation study to previous findings (Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Next, we review the limitations of the simulation study and outline further research needs. Then, we revisit the empirical example and its relation to the simulation study. We will conclude the section with some recommendations for the application of the generalized ANCOVA for multilevel designs with treatment allocation at the cluster-level. A com-

prehensive discussion of the advantages and the problems of the generalized ANCOVA and the distinction between stochastic and fixed covariates will be given in the general discussion in Chapter 6.

5.4.1 Problems of the Appropriate Statistical Models

Overall, the results of the simulation did not clearly favor a single statistical method for estimation of the average causal effect with the generalized ANCOVA in conditionally randomized and quasi-experimental designs with treatment assignment at the cluster-level under the studied conditions. The two most promising methods that yielded unbiased *ACE*-estimators — the multigroup multilevel latent variable model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) and the adjustment procedure of Croon and van Veldhoven (2007) — were both plagued by specific shortcomings that do not allow their unequivocal recommendation for the analysis of average causal effects in non-randomized multilevel designs with treatment assignment at the cluster-level.

Considering the results of the simulation study in total, the multigroup multilevel latent variable model as implemented in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) was the most promising method: It yielded an unbiased *ACE*-estimator and mostly unbiased standard errors. However, under small (and realistic) $ICC(Z)$-values, the standard error estimator clearly overestimated the variability of the *ACE*-estimator leading to conservative significance tests and — potentially — a loss of power. Additionally, the model had severe convergence problems in samples with 20 clusters. Furthermore, the *ACE*-estimator was less efficient than any other method considered and sometimes outperformed in terms of the *MSE* even by methods that had shown a considerable bias in estimation. Similar results have been obtained by Lüdtke et al. (2008) who reported good properties of the multilevel latent variable model in `Mplus` in terms of parameter bias and coverage starting at samples of 50 clusters, but also found a larger empirical standard deviation of the parameter estimator and loss of efficiency compared to the conventional hierarchical linear model.

The adjustment method of Croon and van Veldhoven (2007) also yielded an unbiased *ACE*-estimator in almost all conditions, except for low *ICC*s of the unit-covariate Z and small sample sizes. These results mirror the findings by Lüdtke et al. (2008) who also found a stronger bias of Croon and van Veldhoven's (2007) procedure in these conditions compared to the full information multilevel latent variable model in `Mplus`.

Additionally, the standard error of the *ACE* was underestimated in some conditions – either because of the assumption of fixed predictors when the general linear hypothesis was used or because of the sparse information available for adjusting the empirical cluster means when the standard error was obtained with `lace`. This negative bias of the standard error led to liberal significance tests. In contrast to the multigroup multilevel latent variable model, the adjustment procedure after Croon and van Veldhoven was not hampered by convergence problems. Taken together, these findings suggest that the adjustment method by Croon and van Veldhoven can also not be recommended unequivocally for practical implementations of the generalized ANCOVA. In contrast to the multigroup multilevel latent variable model in `Mplus`, it is not readily available in an all-purpose software package and has to be specifically tailored to fit the analytical problem at hand.

The present simulation extended previous research on the multilevel latent variable model in `Mplus` (Lüdtke et al., 2008) in two ways: (1) A multigroup multilevel latent variable model was used to allow for interactions between the latent between-component Z_b and within-component Z_w and the treatment variable and for different residual variances at the unit- and at the cluster-level. (2) Cross-level interactions between the between-component Z_b, the within-component Z_w were included in the treatment-group specific multilevel structural equation models. The latter step might be partly responsible for the convergence problems in conditions with 20 clusters, since the number of parameters was comparatively large relative to the number of clusters in each treatment group. However, leaving out the cross-level interaction term in a small additional exploratory study did not increase convergence significantly. Previous research on the singlegroup multilevel latent variable model (Lüdtke et al., 2008) has not reported any convergence problems, but only for samples of at least a total of 50 clusters and in models with no cross-level interactions. Thus, it is possible that even the simpler singlegroup multilevel latent variable model might experience convergence problems with smaller samples of clusters. To simplify the model and make estimation more stable in applications, it seems advisable to sequentially test for the necessity of including additional interaction effects and for heterogeneity of the residual variances between treatment groups in the specification of the model.

5.4.2 Limitations of the Simulation Study

As in all simulation studies, the results previously discussed are restricted to the conditions actually included in the simulation design (Skrondal, 2000). Although the independent variables covered a wide range of realistic parameter values, there were nevertheless some notable structural limitations of the design. In the following section, we first discuss the performance of the inappropriate statistical methods and review the properties of the data generation procedure that put them at a disadvantage. We then briefly discuss the omission of the conventional multilevel ANCOVA without interactions from the simulation design. Next, we review stochastic treatment group sizes and variance heterogeneity that did not influence the results of the simulation although both have been found to be influential in previous studies (Kröhne, 2009; Nagengast, 2006). Finally, we will outline further research needs to evaluate the performance of the statistical models.

Inappropriate Statistical Models

The data generation procedure was explicitly modeled after the multilevel single-unit trial introduced in Chapter 2 and consisted of a series of independent repetitions of a stable multilevel random experiment. This stacked the deck against the conventional hierarchical linear model in `nlme` and the singlelevel implementation of the generalized ANCOVA in `lace` in four ways, that should be taken into account when trying to generalize the results of the simulation study:

1. The data were explicitly generated with a multilevel structure of the effects of the unit-covariate Z: The effects of Z_b and Z_w and their product on the true-outcome variable τ_0 and the true-effect variable δ_{10} differed considerably [see also Section 2.3 and Equations (5.37) and (5.38)]. This put the naive implementation of the adjustment model in `lace` at a disadvantage, since it only included regression coefficients for the unit-covariate Z. Predictably, the *ACE*-estimator showed a strong bias. If either the effects and interaction of the within-component Z_w and the between-component Z_b had been the same or if the *ICC* of the unit-covariate Z had been equal to zero, the naive model implementation in `lace` would have yielded an unbiased *ACE*-estimator, but not necessarily the correct standard errors.

2. The data generation procedure also included residual variance components at the cluster-level $Var(r_{0;C})$ and $Var(r_{10;C})$ [see Equations (5.33) and (5.34)] that captured residual effects of the cluster variable C on the true-outcome variable τ_0 and the true-effect variable δ_{10} over and above the influences of the between-component Z_b and the within-component Z_w. This put the implementation of the full adjustment model in `lace` at a disadvantage, since it did not include parameters that would have captured these variance component and was thus likely to yield negatively biased standard errors (Hedges, 2007a; Moerbeek et al., 2000, 2001; Raudenbush, 1997; Snijders & Bosker, 1999). If the two variances $Var(r_{0;C})$ and $Var(r_{10;C})$ had been zero, i.e., if the residual intraclass correlations of the true-outcome variable τ_0 and δ_{10} had been zero, after controlling for all covariates, the singlelevel implementations of the adjustment model in `lace` would have also yielded correct standard errors for the average causal effect (Snijders & Bosker, 1999). The bias of the *ACE*-estimator due to the use of the fallible cluster means of the unit-covariate as predictors would have remained.

3. At data generation, the values of Z_b, i.e., the expected values $E(Z|C{=}c)$, were used as cluster-covariates, while only the fallible cluster means of Z were available as manifest predictors in the data sets. This put all methods at disadvantage that did not explicitly handle the bias in the estimated regression parameters due to the unreliability of the cluster means. Predictably, the model implementations in `lace`, `nlme` and the singlegroup multilevel model in `Mplus` yielded biased *ACE*-estimators in line with the analytical derivations by Lüdtke et al. (2008) and Snijders and Bosker (1999) [see also Equation (2.14)], even though they separately modeled the between-component Z_b and the within-component Z_w by including the cluster-means and the cluster-mean centered values of the covariate as predictors. If the confounding cluster-covariate had not been the latent variable Z_b, but the empirical cluster means or another cluster-covariate V measured without error — as, e.g., in conditionally randomized designs with randomization conditional on the observed values of the between-component or other cluster-covariates — the performance of these models would have been better (see, Lüdtke et al., 2008, for a similar distinction between reflective and formative contextual variables). However, the conventional hierarchical linear model in `nlme` would have still suffered from not taking the stochasticity of the unit-covariate Z into account.

4. Finally, by repeatedly sampling from the single-unit trial, the realized values of the unit-covariate Z varied from sample to sample making Z_b and Z_w stochastic predictors. This put the conventional linear model implementation in `nlme` at a disadvantage, since it explicitly assumes fixed predictors that are constant over replications of the simulation (Pinheiro & Bates, 2000). Consequently and in line with the derivations by Chen (2006) and Kröhne (2009), this model implementation exhibited a negatively biased standard error in conditions with strong interactions and strong dependencies between Z_b and the treatment variable X. An extended discussion of modeling covariates as fixed or stochastic predictors in the specification of the generalized ANCOVA in applications is foregone until the general discussion in Chapter 6.

Conventional Multilevel ANCOVA

Again, the conventional multilevel ANCOVA without interactions between the treatment variable X and the covariates at the unit- and at the cluster-level — usually discussed as a means to heighten precision in randomized designs — (Bloom et al., 1999, 2007; Donner & Klar, 2000; Gitelman, 2005; Moerbeek et al., 2001; Murray, 1998; Oakes, 2004; Raudenbush, 1997; Raudenbush et al., 2007; VanderWeele, 2008) was not tested in the simulation study and compared to the generalized ANCOVA model. Additionally including this model and its implementation in the different statistical frameworks would have made an already complex simulation design even more complicated and would likely not have yielded further insights above and beyond the known analytical and simulation results: An ANCOVA model, erroneously specified without interactions, does not identify the average causal if interactions are present and the conditional effect function $CCE_{jk;V,Z_b}$ is not a constant (Kröhne, 2009; Flory, 2008; Rogosa, 1980). A conventional ANCOVA would have given an adequate estimator of the average causal effect, if no interactions between the treatment variable X and the covariates had been present. While there were conditions without an interaction of the treatment variable X and the between-component Z_b, the interaction of the treatment variable X and the within-component Z_w was constant across the conditions of the simulation design and always different from zero. Hence, unbiased performance of the conventional ANCOVA would not have been expected. Although the robustness of the conventional ANCOVA for non-randomized multilevel designs with treatment

assignment at the cluster-level (depending on its implementations in statistical models) remains a topic for further research, the use of the generalized ANCOVA, that encompasses the conventional ANCOVA without interactions as a special case and always correctly identifies the average causal effect, is recommended.

Stochastic Group Sizes and Variance Heterogeneity

In contrast to previous findings in simulation studies of the singlelevel generalized ANCOVA implemented in multigroup structural equation models (Kröhne, 2009; Nagengast, 2006), it was not necessary to treat the sizes of treatment and control group as random variables. Even though the implementation of the adjustment model in the multigroup multilevel latent variable model in `Mplus` had to compute not only the estimator of the expected value $E(Z)$, but also of the expected value $E(Z \cdot Z_b)$ and various variance parameters from treatment-group specific parameters (see Appendix A.2) — using the treatment group sizes to this effect — and the size of the treatment groups varied between replications, the corresponding standard error did not underestimate the empirical variability of the *ACE*-estimator markedly. Most likely, this was due to the smaller effect sizes of the interaction considered in the present study. A noticeable advantage for implementations of the multivariate delta method with random group sizes has been previously shown under comparatively stronger effect sizes of the interaction (see Kröhne, 2009; Nagengast, 2006). Additionally, only equally-sized treatment and control groups (in terms of clusters and absolute number of units) were considered. Nevertheless, the augmented covariance matrix of the model parameters proposed by Nagengast (2006) could be implemented for the multigroup multilevel latent variable model in `Mplus` to calculate standard errors that take stochastic group sizes into account. An open question for this implementation is the choice of the appropriate treatment group size, since it would be possible to consider either the number of clusters per treatment condition or the number of units in this regard.

Although the residuals $\upsilon_{10;U}$ and $r_{10;C}$ of the conditional causal effect function had variances larger than zero [$Var(\upsilon_{10;U}) > 0$ and $Var(r_{10;C}) > 0$], resulting in slightly heterogeneous conditional variances of the outcome variable Y in treatment and control condition on the unit- and the cluster-level, they did not influence the simulation results markedly. The bias in the standard error of the `nlme` implementation, that did not take this heterogeneity into account, was due to treating predictors as fixed not due to

omission of an additional variance component — a small exploratory simulation study that allowed for heterogeneous unit-level variances in `nlme` yielded almost identically biased standard errors. In a similar vein, the standard errors of the singlegroup multilevel manifest variable model in `Mplus` yielded unbiased standard errors, although they did not take the variance heterogeneity into account. These findings are in line with Korendijk et al. (2008) who also found no effects of misspecified variance heterogeneity on the cluster-level on the fixed effects in a conventional hierarchical linear model. Nevertheless, results from simulation studies of the singlelevel generalized ANCOVA (Kröhne, 2009) indicate that larger variance heterogeneity between the treatment groups and unequal treatment group sizes could make parameter estimates and standard errors inconsistent, if they are obtained from implementations that do not take this heterogeneity explicitly into account. This would put the statistical models that explicitly model this heterogeneity, such as the multigroup multilevel latent variable model in `Mplus`, at a distinct advantage. The amount and consequences of a larger variance heterogeneity at both the unit- and the cluster-level remain open questions worthy of further studies.

Further Research Needs

Finally, two especially glaring omissions from the simulation design need to be addressed in further simulation studies: First, only a single unit-covariate Z decomposable into its between-component Z_b and its within-component Z_w was included in the data generation. Although the simulation study demonstrated the considerable complexities of estimating the *ACE* with the generalized ANCOVA even in this simple constellation, it does not speak to the additional complexities and sample size requirements involved in specifying and estimating a model with more than one covariate at the unit- or at the cluster-level. In this case, the correct specification of interactions between the covariates becomes a critical and complicated issue. This is problematic insofar as it is unlikely that controlling a univariate covariate at the unit-level and the cluster-level will ever suffice to achieve unbiasedness of $E(Y \mid V, Z_b)$ in quasi-experimental multilevel designs with treatment assignment at the cluster-level. However, in designs with conditional randomization of clusters to treatment conditions, randomized assignment conditional on a single cluster-covariate is possible.

A second major shortcoming of the present simulation study is the fact that the implementations of the generalized ANCOVA were only compared under the null hypoth-

esis of no average causal effect. While the correct estimation of parameters and their standard errors under the null hypothesis are important for every statistical model to guarantee appropriate tests of statistical significance, they are not sufficient for final conclusions about the applicability and usefulness of a statistical procedure. Especially in the planning of evaluation studies, the power of a design and the statistical analysis for detecting a treatment effect of a certain magnitude is of major interest (Moerbeek et al., 2000, 2001; Murray, 1998; Raudenbush, 1997). The present study speaks to these issues only insofar, as those methods that yield biased *ACE*-estimators and standard errors even under the null hypothesis are clearly not recommendable. The modest efficiency of the multigroup multilevel latent variable model compared to other procedures hints at a likely loss of power when this method is used to estimate and test average treatment effects in applications. However, the amount of this potential loss and its tradeoffs with a biased parameter estimator remain to be studied.

5.4.3 Example Analysis

The example analysis of the Early Childhood Longitudinal Study (ECLS-K, National Center for Education Statistics, 2001) data set was intended to illustrate the performance of various implementations of the generalized ANCOVA with an empirical example that was structurally similar to the simulation study. Although the results of these analyses — no average effect of a mud-play area on quantitative competencies after controlling for the between- and within-components of the pre-test — cannot be interpreted causally without the additional untested assumption of conditional unbiasedness. It is likely that other cluster-covariates V or between-components Z_b could influence both the outcome variable and the assignment of the treatment over and above the cluster-means of the pre-test. Nevertheless, the analysis was illustrative of the implementations of the adjustment models and the complexities involved in interpreting effect estimates from different statistical models in practice.

The different statistical models behaved more or less as expected compared to similar conditions of the simulation design: While the numerical *ACE*-estimates — except for the multigroup multilevel latent variable model in `Mplus` — were roughly equal after transforming them to a common effect size metric, the standard errors differed considerably, in line with the predictions: The standard errors from `lace` were smaller than the comparable estimates from models that included the appropriate variance components.

Due to the small effect size of the interaction between X and Z_b, the standard error of the `Mplus` implementation was only slightly larger than the standard error obtained from `nlme`. Nevertheless, the statistical inferences from the different implementations were similar — except for the multigroup multilevel latent variable model in `Mplus` and (unsurprisingly) the full adjustment model in `lace` that were the only models that indicated a statistically significant average treatment effect, albeit a very small one.

In light of the simulation study and the presumed data constellation in the empirical example, this was not surprising: Since the empirical cluster means of the pre-test were not computed from all students within the kindergarten class, but only from those included in the ECLS-K sample, they were fallible measures of the corresponding values of the between-component Z_b. Hence, the average effect estimator was likely to be positively biased under the parameter constellation, if this unreliability was not corrected. Nevertheless, the different statistical (though not necessarily substantive) conclusions about the average treatment effect illustrate the problem of generalizing from a simulation study to applied contexts: Although the simulation showed that the multigroup multilevel model in `Mplus` was appropriate under the null hypothesis of no average causal effect, the true average causal effect is not known in applications and the results of the simulation do not automatically generalize to these conditions.

5.4.4 Recommendations and Conclusion

Considering the results of the simulation as a whole, the outlook for the use of the adjustment model in applications is not too bright and no method can be unequivocally recommend for applications: Available sample sizes which are typically in the range of 20 to 50 clusters may not suffice for stable implementations of the most successful implementation of the full adjustment model. The implementation of the full adjustment model in the multigroup multilevel latent variable model in `Mplus` did not converge reliably in these conditions and showed a significant positive bias of the standard error under small $ICC(Z)$-values. The adjustment procedure of Croon and van Veldhoven (2007) showed better convergence properties, but the standard error underestimated the empirical variability of the ACE-estimator markedly with small cluster sizes and a small number of clusters. Conventional implementations of the hierarchical linear model and the implementation of the singlelevel generalized ANCOVA in `lace` were appropriate only for a subset of conditions — uncritically using them in applications cannot

be recommended for multilevel designs with treatment assignment at the cluster-level. Alternative implementations of the generalized ANCOVA model such as corrections of the unreliability of the cluster-means developed by Grilli and Rampichini (2008) and direct corrections of standard errors for clustering (e.g., Hedges, 2007a) may be alternatives that need to be considered in further studies.

Until then, the multigroup multilevel model implementation of the generalized ANCOVA model for non-randomized designs with treatment assignment at the cluster-level can be cautiously recommended as the model of choice, if the latent between-component Z_b influences the outcome variable Y and the treatment assignment probabilities: If the number of clusters is sufficiently large, i.e., not smaller than 50, the model yielded an unbiased *ACE*-estimator. It should be taken into account, that the standard error obtained with this model might be overestimated under small intraclass correlations of the unit-covariate. Additionally and to enhance the stability of model estimation, careful specification tests with respect to variance components and interactions are highly recommended.

While we have focused on the statistical implementations and the simulation study in the preceding section, the general appropriateness of the generalized ANCOVA for causal inferences and other overarching problems will be given further attention in the general discussion in Chapter 6.

6 General Discussion

In the general discussion, we review the theoretical derivations, simulation results and example analyses presented in this thesis and critically discuss limitations and open questions. The discussion follows the structure of the thesis: We first review the theory of causal effects and point out some limitations of its scope and breadth. Then, we review the generalized ANCOVA, discuss critical assumptions and its relation to other procedures for identifying and estimating average causal effects. We also return to the properties of the repeated single-unit trials in the simulation studies and will review the ensuing consequences for the statistical models and their appropriateness. We conclude with an outlook on further research needs for causal inference in multilevel designs.

6.1 Causal Inference in Multilevel Designs

We started this thesis noting that causal inference in multilevel designs has been relatively little studied, although multilevel designs potentially pose additional challenges for the definition of causal effects and their proper analysis. While a broad statistical literature deals with the analysis of randomized designs (e.g., Donner & Klar, 2000; Moerbeek et al., 2000, 2001; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000), an explicit foundation of the statistical procedures in a theory of causal effects is widely lacking — this is especially critical for quasi-experimental designs. Previous applications of Rubin's (1974, 1977, 1978) theory of causal effects to multilevel designs were often restricted to specific problems or case studies, inadequately formalized and made contradictory claims with regard to the conditions and assumptions that have to be fulfilled for valid causal inferences. In this thesis, we sought to fill this gap and developed a theory of causal effects in multilevel designs that captures their peculiarities in a rigorously formalized framework. We reconciled and clarified inconsistencies between existing accounts of causality in multilevel designs. Finally we developed, tested

and applied procedures for the analysis of average causal effects based upon this theory.

In this section, we will first review the main results concerning the general theory of causal effects and its application to multilevel designs. We will then critically discuss the interpretation of the treatment variable X, the cluster variable C and the multilevel single-unit trial as representation of the empirical phenomena studied in multilevel designs.

6.1.1 Review

In Chapter 2, we outlined the general theory of causal effects (Steyer et al., 2009) and showed how it can be used to represent multilevel designs. Based upon two single-unit trials for different classes of multilevel designs, we introduced a causality space consisting of a probability space, the putative cause X that is pre-ordered to the outcome variable Y, a filtration of sub-σ-algebras that represents the time-order of events and a confounder σ-algebra $\mathfrak{C}_X$. The causality space and the confounder σ-algebra $\mathfrak{C}_X$ both explicitly included the cluster variable C and cluster-covariates V. We showed how core concepts of multilevel analysis (e.g., the intraclass correlation coefficient) are defined with respect to the distributions of events and variables of the multilevel causality space. We then defined average causal effects as well as conditional causal effects and showed how individual causal effects (Neyman, 1923/1990; Rubin, 1974, 1977, 1978) and cluster-specific individual causal effects (Gitelman, 2005) are represented within the general theory of causal effects. Next, we introduced unbiasedness as the weakest causality criterion for the identification of average and conditional causal effects with the empirically estimable prima-facie effects. Finally, we introduced sufficient conditions for unbiasedness and discussed their relevance for applications to multilevel designs in practice.

In Chapter 3, we refined our analysis of causal inference in multilevel designs and studied violations of the *Stable Unit Treatment Value Assumption* (SUTVA, Rubin, 1977, 1986, 1990). Such violations are discussed as one threat to the validity of effect definitions and the meaningful interpretation of average causal effect estimates from samples in multilevel designs (Gitelman, 2005; Hong & Raudenbush, 2006, 2008; Oakes, 2004; Sobel, 2006). We compared different alternative stability assumptions and analyzed how their appropriateness depends on the theoretical status of the cluster variable C and on assumptions about the assignment process of units to clusters. Further-

more, we showed that, within the general theory of causal effects, the effect definitions are not invalidated by confounding effects of the cluster variable or interferences between units that are captured in variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$. Meaningful estimation of causal effects from samples, however, requires repetitions of independent single-unit trials that are invariant with respect to the causal parameters and distributions. In the second part of Chapter 3, we systematically reviewed different types of between-group multilevel designs and classified them along three dimensions — assignment of units to clusters, level of treatment assignment and treatment assignment mechanism — to establish a taxonomy of multilevel between-group designs.

6.1.2 Interpretation of Variables

The theory of causal effects offers, first and foremost, a mathematically formalized representation for the study of causal relations in general and for causal effects in multilevel designs, as studied in this thesis, in particular. Apart from the temporal structure and other assumptions made in the definition of the causality space, it does not speak to the substantive interpretation and meaning of the random variables. The theory does, however, precisely outline the conditions under which unbiased estimators of the average causal effect can be obtained. In terms of the Campbellian tradition (Campbell & Stanley, 1966; Cook & Campbell, 1979; Shadish et al., 2002) of experimental and quasi-experimental design, it is mostly concerned with the internal validity of a design. Nevertheless, issues of external validity (Shadish et al., 2002) are also important if substantial inferences are to be drawn from multilevel designs: Especially the interpretation of the treatment variable X, the cluster variable C and the between-component Z_b deserve special attention, as does the multilevel single-unit trial as a representation of the empirical phenomenon that is studied.

Treatment Variable

In the introduction of the treatment variable X, we only assumed that different treatment and control conditions existed and that they could be implemented in all clusters. Differences in the treatment group-specific and covariate-conditional expectations of the outcome variable Y due to the treatment variable X were used to define causal treatment effects, without further assumptions about the processes to bring these differences

about. In reality, treatment conditions usually differ in more than one respect (e.g., not only in the actual medical treatment a patient receives, but also in the attention given to the patient by the medical staff, the frequency of visits to the care provider, etc.). Which of these differences between treatment conditions are causally relevant for bringing differences in the outcome variable about might not be instantiously obvious. In order to pinpoint such causal mechanisms, careful planning of evaluation studies based on insights into substantive matters is necessary (e.g., by holding constant some aspects of treatment implementation across treatment conditions and by implementing a rigorous treatment protocol, see, Shadish et al., 2002). The theory of causal effects is indifferent to the substantive interpretation of the treatment variable, it only outlines conditions under which causal effects can be attributed to the treatment as such; the proper substantive interpretability has to be guaranteed by carefully choosing the conditions to be compared.

Another issue that threatens the correct interpretation of treatment effects specifically in multilevel designs are varying implementations of the treatment in different clusters. In our formal introduction of causal effects, we defined the effects of treatment variable X, assuming that the treatment could be implemented similarly in all clusters. However, in the practice of multisite evaluation, fidelity to the treatment protocol is a strong concern and variations in treatment implementation between clusters are common (e.g., Turpin & Sinacore, 1991). Including factors that influence treatment implementation as cluster-covariates (Seltzer, 2004) is one possible remedy for differences in treatment implementation. Doing so guarantees that the average causal effect estimator will at least be unbiased with regard to these covariates. On the other hand, it still only allows inferences about the treatment as it was implemented in practice; not about the hypothetical ideal implementation of the treatment in all clusters. If such inferences are desired, the solution cannot come from theoretical or statistical models, but must rely on careful design and controlled implementation of the evaluation study.

Cluster Variable and Cluster-Covariates

A similar concern applies to the interpretation of the cluster variable C and covariates at the cluster-level, especially the between-component Z_b. The theory of causal effects treats the cluster variable C like any other random variable measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$. Substantively, it can have many different interpretations:

The locations of neighborhoods (Sobel, 2006), administrative practices in schools (Raudenbush & Willms, 1995), environmental conditions (Oakes, 2004), effects of different therapists (Gitelman, 2005), interferences between units as well compositional effects (in designs with pre-exisiting clusters, Hong & Raudenbush, 2006; Sobel, 2006) and many more are captured by the cluster variable C. The mechanisms by which the cluster variable C can influence the outcome, interact with the treatment and confound the relation between treatment and outcome are manifold. Similarly, the interpretation of the between-component Z_b and the mechanisms by which it influences the outcome variable differ from context to context. Such mechanisms may be located within the individual unit (e.g., may be brought about by comparisons with other students in a school, Marsh, Hau, & Craven, 2004) or at the level of the cluster (e.g., by providing an especially motivating working climate, Van Mierlo, Vermunt, & Rutte, 2009). The general theory of causal effects subsumes all of these mechanisms under the definition of the cluster variable C and the between-component Z_b. However, it does not — at least as presented here — claim that such effects can be interpreted causally. The cluster variable C and the between-component Z_b are simply considered as variables measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$, used in the definition of the true-outcome variables τ_j and as covariates in the identification and estimation of average causal effects — evaluating their effects substantively has its own merits and research tradition in the analysis of contextual effects (e.g., Cronbach, 1976; Croon & van Veldhoven, 2007; Kozlowski & Klein, 2000) and institutional comparisons and value-added models (e.g., Fiege, 2007; McCaffrey et al., 2004; Raudenbush & Willms, 1995; Wegscheider, 2004).

Multilevel Random Experiment

The substantive meaning of the treatment variable X, the cluster variable C and the between-component Z_b are also directly relevant for the adequacy of the multilevel random experiments and single-unit trials: The two classes of multilevel single-unit trials are intended to capture the respective structure of the design under consideration; all inferences are restricted to the distributions of events in this random experiment (Steyer et al., 2009). This restricts the set of meaningful generalizations: A random experiment always refers to a specific set of units and clusters, specific sampling probabilities of units and clusters, a specific assignment mechanism and the temporal order of assignment of units to clusters, time-lags between the selection of units and clusters, registration of

covariates and the onset of the treatment and so forth. Every evaluation study designed to estimate causal effects of a treatment has to meticulously specify to which single-unit trial it refers. The two general classes of single-unit trials for multilevel designs introduced in Chapter 2 are no more than starting points in this regard. Applications may require the representation of different temporal sequences of the assignment of units to clusters and the assessment of of covariates at both the unit- and the cluster-level to adequately represent all aspects of the considered design. The consequences for the validity of effect definitions must be carefully scrutinized. Further generalizations to side conditions, such as historical context or geographic location, that are not explicitly represented in the single-unit trial, require additional assumptions. Causal effects are therefore always restricted — in their definition — to the single-unit trial under consideration; their estimation from a sample always refers to repetitions of the multilevel single-unit trial that are invariant with respect to the causal parameters and distributions. Inferences and conclusions about other timepoints, locations or different boundary conditions are not backed by the random experiment and must rely on substantive arguments and assumptions about the similarities between the two single-unit trials considered.

6.2 Generalized ANCOVA

In the following section, we will first review the generalized ANCOVA for multilevel designs. We will then discuss (1) the problems of its implementation in statistical models, (2) conceptual problems inherent in linear models of the outcome variable, (3) alternative adjustment procedures and (4) the critical assumption of conditional unbiasedness.

6.2.1 Review

In Chapter 4, we built upon the theory of causal effects to adapt the generalized ANCOVA (Steyer et al., 2009) to conditionally randomized and quasi-experimental designs with treatment assignment at the unit-level. We derived the generalized ANCOVA for designs with an unbiased unit-covariate-cluster-treatment regression $E(Y \mid X, Z, C)$ and for designs with an unbiased unit-covariate-cluster-covariate-treatment regression $E(Y|X, Z, V, Z_b)$. For each design type, we showed how the average causal effect is iden-

tified generally and under the assumption of linear effect and intercept functions. We extended previous accounts of multilevel ANCOVA by identifying the average causal effect in models with interactions between the treatment and the covariates. We then studied several implementations of the generalized ANCOVA for designs with unbiasedness of $E(Y \mid X, Z, V, Z_b)$ using a data generation procedure consisting of identical and independent repeated multilevel single-unit trials. In line with results from simulation studies for singlelevel designs, we showed that the stochasticity of the covariates has to be taken into account in the computation of standard errors for the *ACE*-estimator in the presence of interactions, rendering the conventional hierarchical linear model that assumes fixed predictors inadequate. The singlegroup multilevel structural equation model in `Mplus` provided an unbiased standard error, but was hampered by convergence problems and required large samples of clusters. Finally, we illustrated different implementations of the generalized ANCOVA with an empirical example from the National Educational Longitudinal Study - Class of 1988 (NELS:1988, Curtin et al., 2002) and discussed the specifics of the simulation study in detail.

In Chapter 5, we developed the generalized ANCOVA for conditionally randomized and quasi-experimental multilevel designs with treatment assignment at the cluster-level. We showed how the average causal effect could be identified under the assumption of unbiasedness of the cluster-covariate-treatment regression $E(Y \mid X, V, Z_b)$ generally and under the additional assumption of linear intercept and effect functions. Once again, we extended previous accounts of multilevel ANCOVA by unambiguously identifying average causal effects in models with interactions between the treatment and the covariates. We showed that two versions of the adjustment model — either using only the between-component Z_b or both the between-component Z_b and the within-component Z_w — yielded identical *ACE*-estimators. Once again, we compared different implementations of the adjustment models in a large simulation study using a data generation procedure consisting of identical and independent repeated multilevel single-unit trials. The results showed that methods that assume fixed predictors yielded biased standard errors. A second important finding was the need to model that empirical cluster means as fallible measures of the corresponding regression $E(Z \mid C)$: Only methods that accounted for the unreliability of the empirical cluster means, specifically the adjustment procedure of Croon and van Veldhoven (2007) and the multigroup multilevel latent variable model in `Mplus` 5.0 (Lüdtke et al., 2008; L. K. Muthén & Muthén, 1998-2007) yielded an unbiased *ACE*-estimator. However, a relatively large number of

clusters was required for adequate convergence of the latter model. We illustrated different implementations of the generalized ANCOVA with an empirical example from the Early Childhood Longitudinal Study - Kindergarten Class of 1998-99 (ECLS-K, National Center for Education Statistics, 2001) and discussed the specifics of the simulation study and its limitations in detail.

In both cases, the generalized ANCOVA studied theoretically and in the simulation studies extended the ANCOVA conventionally discussed for multilevel designs: It provided for non-constant conditional effect-functions that could vary with the covariates, i.e., it included interactions between the covariates and the treatment. Previous applications of the conventional multilevel ANCOVA to multilevel between-group designs had either not included interactions at all (Bloom et al., 1999, 2007; Moerbeek et al., 2001; Oakes, 2004; Raudenbush, 1997; Raudenbush & Liu, 2000) or had not provided a straightforward definition of the average causal effect in presence of interactions (Gitelman, 2005; Pituch, 2001; Plewis & Hurry, 1998; Seltzer, 2004; VanderWeele, 2008). The conventional ANCOVA — usually understood as a model that only includes the main effects of the predictors (see Kröhne, 2009, for a comprehensive review) — does not identify the average causal effect in the presence of covariate-treatment interactions, but only the conditional treatment effect at the point of highest precision (Rogosa, 1980). Hence, it is only appropriate, if the conditional effect function is a constant and there are no treatment covariate interactions. The generalized ANCOVA, on the other hand, correctly identifies the average causal effect in both the presence and absence of treatment-covariate interactions, includes the conventional ANCOVA without interactions as a special case and covers all possible eventualities in applications.

6.2.2 Multilevel Models

Both simulations showed that the structure of the multilevel single-unit trial influenced the specification of the generalized ANCOVA and its estimation in statistical procedures in two ways: (1) Conceptually, the decomposition of the unit-covariate into the between-component Z_b and the within-component Z_w had to be modeled to obtain unbiased *ACE*-estimators. (2) Statistically, variance components that accounted for the effects of the cluster variable had to be included in the models to obtain unbiased standard errors.

Multilevel Decomposition of the Unit-Covariate

Both simulation studies used a data generation procedure in which the between-component Z_b and the within-component Z_w influenced the true-outcome variables τ_j in the treatment and the control group independently. This accounted for contextual effects of the unit-covariate brought about by different compositions of the clusters and extended some of the existing accounts of multilevel ANCOVA that do not include contextual effects (Donner & Klar, 2000; Gitelman, 2005; Murray, 1998; Pituch, 2001; Plewis & Hurry, 1998). As expected from theoretical derivations, models that did not take this decomposition into account, yielded biased *ACE*-estimators in both simulation studies. This highlights the conceptual need to take the multilevel structure of the effects of the unit-covariate into account when analyzing between-group multilevel designs with the generalized ANCOVA. Nevertheless, there are situations, in which using only the unit-covariate Z as predictor will lead to an unbiased *ACE*-estimator: If the *ICC* of the unit-covariate Z is equal to zero or if the effects of the between-component Z_b and the within-component Z_w on the true-outcome variables τ_j are equal in all treatment groups, the two components do not have to modeled separately. However, these conditions were not explicitly included in the simulations in Chapter 4 and 5. Therefore, the simulation studies do not speak to the convergence behavior and appropriateness of statistical methods in these conditions. In applications, significance tests of the *ICC* of the unit-covariate (see Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) and tests for the difference of the respective regression weights (Enders & Tofighi, 2007; Kreft et al., 1995) are recommended to obtain a parsimonious statistical models implementation of the generalized ANCOVA.

Whether or not the between-component Z_b and the within-component Z_w have to be modeled as latent variables depends on the actual properties of the sample, such as the average cluster sizes and the intraclass correlation coefficient of the unit-covariate, but also on the nature of the contextual covariates considered. If the true values of Z_b and Z_w influence the treatment assignment probabilities and the outcome variable (and they cannot be reliably estimated in practice), it is advisable to explicitly account for this unreliability. If Z_b can be reliably estimated, e.g, if *all* units within a cluster are assessed and the unit-covariate Z is measured without error, it is not necessary (and might be even detrimental) to explicitly model Z_b as a latent variables (see, Lüdtke et al., 2008). The same holds, if the observed cluster means of the unit-covariate Z are

used to randomize units or clusters to treatment conditions. In this case, the appropriate covariate with regard to which unbiasedness holds is not the latent between-component Z_b, but the manifest, fallible cluster means. Adjustment models have to be specified accordingly.

Variance Components

Both simulation studies used a data generation procedure without conditional homogeneity of the true-outcome variables τ_j with regard to the unit-covariate Z and the between-component Z_b. In fact, both the unit-variable U and — more importantly — the cluster variable C influenced the true-outcome variables τ_j after accounting for the covariates. As expected, singlelevel adjustment models that did not include the proper variance components to account for the lack of conditional homogeneity of the true-outcome variables due to the cluster variable C underestimated the variability of the *ACE*-estimator in both simulation studies. This highlights the need to take the multilevel structure into account when analyzing data from conditionally randomized and quasi-experimental multilevel designs and not to use singlelevel adjustment procedures uncritically. Nevertheless, there are situations, in which singlelevel models yield correct standard errors: This will be the case, if the residual intraclass correlation coefficients (*rICC*) of the true-outcome variables τ_j will be equal to zero after controlling for all covariates at the unit- and at the cluster-level (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). As already discussed in more detail in Chapters 4 and 5, these conditions were not included in the simulation designs — the *rICC* and the corresponding variance components were held constant across conditions. Therefore, the results do not speak to the convergence behavior and the relative efficiency of multilevel and singlelevel models in these situations. However, multilevel models such as the multilevel latent variable model (Lüdtke et al., 2008) were the only models to yield unbiased *ACE*-estimators if the between-component Z_b confounded the average treatment effects and are naturally suited to statistically test the residual *ICC* of the outcome variable Y.

6.2.3 Stochastic Predictors

The second consistent finding in both simulations studies referred to the nature of covariates in quasi-experimental and other multilevel designs: The data generation procedures, and the multilevel single-unit trials they reflected, resulted in covariates that

were random variables whose realized values varied from sample to sample. The results indicated that the assumption of fixed predictors conventionally made in the general linear model (Searle, 1971; Kutner et al., 2005; Rechner & Schaalje, 2007; Werner, 1997) or the hierarchical linear model (Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002) leads to negatively biased standard errors of the average causal effect estimator if the predictors are stochastic, i.e., if the realized values of the covariates vary from sample to sample. While the differentiation in models with fixed and stochastic predictors does not influence the parameter estimators, omnibus model tests and confidence limits for regression weights in the general linear model (Sampson, 1974), it becomes relevant when interactions between predictors are considered among the set of linear predictors (and also for the construction of confidence limits for the coefficient of determination, see Algina, 1999; Steiger & Fouladi, 1997). In this case, the assumption of multivariate normality of the predictors, on which the equivalence proofs of linear regression models with fixed and stochastic regressors rely, is violated (Fisicaro & Tisak, 1994) and the computation of standard errors has to take the additional variability induced by the stochastic nature of the covariates into account. Analytical derivations (Chen, 2006; Kröhne, 2009) and simulation studies (Flory, 2008; Kröhne, 2009; Nagengast, 2006) have consistently shown that the distinction between fixed and stochastic predictors is relevant for the estimation of the standard error of the average causal effect and significance tests of the *ACE*-estimator obtained with the generalized ANCOVA. This is not surprising since the distribution of the regression coefficients differ between the model with fixed and the model with stochastic predictors (Sampson, 1974) and the mean of the covariate that is used in the non-linear constraint is only a fallible estimate of the expected value of the covariate in the population.

Stochastic predictor variables naturally occur, if one assumes that samples are generated by repeating the single-unit trial introduced in Chapter 2. This is especially relevant for quasi-experimental designs and observational studies in applications (see also, Chen, 2006; Gatsonis & Sampson, 1989; Schafer & Kang, 2008): Such designs will almost always result in different realizations of covariates and other variables in the sample. Thus, it is not appropriate to consider the sample realizations of the variables as fixed. Correct inferences about the underlying distributions require that predictors are modeled as stochastic.

6.2.4 Challenges to the Generalized ANCOVA

The generalized ANCOVA, although theoretically an adequate procedure and at least promising in simulation studies, has to confront a number of specific problems and challenges in applications. Those are (1) the correct specification of intercept and effect functions, (2) extrapolation beyond the observed scores of the covariates in a sample and (3) the relevance of average causal effects in the presence of interactions.

Specification of Effect Functions

The generalized ANCOVA — for singlelevel as well as for multilevel designs — relies on correctly specified intercept and effect functions. Specifically, both functions are assumed to be linear in the covariates and — as far as they are also considered — their product variables. This assumption makes the *ACE*-estimator of the generalized ANCOVA vulnerable to non-linearities: If the effect functions are misspecified, their expected values will no longer be equal to the *ACE*. Thus, specification tests become critical in practical applications: If all covariates are discrete random variables (and the sample is sufficiently large), it is possible to compare the conditionally linear regression model to a saturated model obtained by representing each combination of covariate values with an indicator variable. While such test can be easily implemented for singlelevel designs (Steyer, 2002), they have not yet been developed and tested for multilevel models. Additionally, formal tests of the linearity assumption are not possible when continuous covariates are considered. In order to informally check the appropriateness of the linearity assumption, the different properties of the residual of the true conditionally linear regression as compared to conditionally linear ordinary least-squares regression can be used: The residual of a true regression is regressively independent of its regression; this is not true of the residual of a ordinary least-squares regression (Kutner et al., 2005; Rechner & Schaalje, 2007). If violations of the regressive independence of the residual from the regressors are detected either by visual inspection for obvious violations or by statistical tests of association, the regression is misspecified (see also Raudenbush & Bryk, 2002; Snijders & Bosker, 1999, for similar tests of the specification of the hierarchical linear model). In order to correctly identify the *ACE* with the expected value of the effect function, a different functional form has to be chosen.

A related problem is the correct specification of the intercept and effect function

when more than one covariate is considered. In the adjustment models for linear effect functions developed in Chapters 4 and 5, all interactions between the covariates were included. While this approach is feasible as long as the number of covariates is small, it becomes quickly impractical with more than two covariates, because the number of potential interaction terms grows exponentially with the number of covariates. However, the need for a parsimonious model specification must be carefully balanced with the requirement to include all relevant covariates that are required for conditional unbiasedness and for the correct specification of the effect function in order to obtain an unbiased *ACE*-estimator.

Extrapolation

A second problem that applies to the generalized ANCOVA for singlelevel as well as for multilevel designs is the extrapolation of results over the range of observed data. This problem of data sparseness or lack of common support (Lechner, 2001; Oakes, 2004) is especially critical in multilevel designs with their predominantly small samples of clusters: If the probabilities of observing values of some covariates in some clusters are small and only a small sample of clusters is available to estimate the relations between the cluster-covariates and the outcome variable, inferences based on linear models often have to rely on extrapolations beyond the range of observed data. This problem is most likely to occur in designs that use pre-existing clusters: For example, if neighborhoods differ markedly with regard to a covariate such as socio-economic status (Oakes, 2004), in case of hospitals that treat patients populations that vary in symptom severity (Wegscheider, 2004) or in schools that educate widely differing student populations (Fiege, 2007). Extrapolation above the range of observed data is unproblematic if the linearity assumption with regard to the effect function holds and each unit has a non-zero probability of being assigned to the treatment and the control group. In this case, extrapolation does not threaten the causal interpretation of the *ACE*-estimator. If linearity does not hold outside the range of observed data, however, extrapolation is not justified and — even more critically — its appropriateness cannot be empirically assessed.

Relevance of Average Effects.

In accounts of moderated multiple regression (e.g., Aiken & West, 1996; Preacher, Curran, & Bauer, 2006; West, Aiken, & Krull, 1996) it has been argued that the main

effect of a variable is no longer very informative or useful when interactions between predictors are present, since it only represents the conditional effect of the treatment at the point where all other predictors have a value of zero. Similar points have been with respect to hierarchical linear regression models (D. J. Bauer & Curran, 2005). Alternatively, a careful analysis of the regressive dependencies between outcome and predictor variables at different values of the other predictors is suggested (D. J. Bauer & Curran, 2005; Preacher et al., 2006; Rogosa, 1980). In the context of the analysis of causal effects from between-group designs, following these suggestions would amount to the analysis and inspection of conditional causal effects given values of the respective covariates (Steyer et al., 2009). Although the focus of this thesis has been on the identification and estimation of the average causal effect and the specification of the conditional effect function has only been used as a means to obtain this identification, conditional causal effects are important in their own right: They are, for example, informative about differential indications of a treatment, if they vary with covariates at the unit-level. In the context of multilevel designs, they can additionally point to the influence of contextual variables on the treatment effect, if covariates at the cluster-level moderate the treatment effect. Even more important, conditional causal effects in different clusters (C=c) (e.g., Seltzer, 2004) point to differences in the average treatment effects in different clusters. Nevertheless the average causal effect remains a useful concept to be considered for theoretical and practical reasons even in the presence of interactions between covariates and the treatment variable: It still retains its meaning as the expected value of the true-effect variable (Steyer et al., 2009) and is the quantity that is estimated by alternative adjustment procedures (see below). New approaches to the analysis of moderated multiple regression models (Gelman & Pardoe, 2007) also use the concept of averaging over the distribution of other covariates to obtain a measure of the influence of a specific predictor on the outcome. The average causal effects always captures the net effect of the treatment averaged over site-specific properties and the distribution of all covariates. As such, it remains a policy-relevant quantity, e.g., for decisions about the effectiveness of a treatment when the complete population is considered.

6.2.5 Alternative Adjustment Methods

A whole class of alternatives to the generalized ANCOVA has not been discussed in detail in this thesis. While the generalized ANCOVA relies on the specification of the regression of the outcome variable on the covariates and the treatment variable, methods that use the estimated propensity scores such as stratification, matching or weighting (e.g., Rosenbaum & Rubin, 1983, 1984, 1985) model the relation of the treatment assignment probabilities and the covariates. Theoretically, modeling the propensity score has the advantage of combining all information about treatment assignment in the covariates into a single index and of allowing the researcher to specify the adjustment model without knowledge of the outcome variable (e.g., Rubin, 2001).

At first glance, propensity score methods seem to solve the problems of specification of effect functions and overlap in the covariate distributions in treatment and control group. However, little is known with regard to the robustness of propensity score methods to misspecifications (Kang & Schafer, 2006; Schafer & Kang, 2008). Since the estimation of propensity scores also relies on a regression model, they are at least theoretically not immune against misspecifications of this function. The balance criterion that is often used to test the appropriateness of the propensity score model (Rubin, 2005), does not test a sufficient condition for unbiasedness nor is it a formal specification test (Steyer et al., 2009). On the other hand, the problem of missing overlap between treatment and control group in the covariate distribution is explicitly considered in methods that use propensity scores: By combining all information of the influence of the covariates on the treatment assignment probabilities in a single index, propensity scores enable the researcher to investigate the amount of overlap of the propensity score distribution between treatment and control group. Inferences about average causal effects can then be restricted to the areas where overlap between control and treatment group exists (Morgan & Winship, 2007).

So far, propensity score methods have not been studied in detail with respect to their applicability to multilevel designs, but have been used rather uncritically for applications in case studies (Hong & Raudenbush, 2006). In light of the findings of the simulation studies that revealed considerable biases not only in the estimated variability of the *ACE*-estimator of the generalized ANCOVA, but also in the estimator itself when conventional models were applied uncritically to quasi-experimental multilevel designs, adjustment methods based on the propensity score, weighting and matching need to

be studied with respect to necessary modifications for these designs. This seems especially critical with respect to the proper modeling of contextual effects of unit-covariates (Croon & van Veldhoven, 2007; Lüdtke et al., 2008).

6.2.6 Tests of Unbiasedness

Finally, the most important condition for interpreting the average effect estimator obtained from the generalized ANCOVA model (and other adjustment procedures) causally — conditional unbiasedness — deserves some specific elaboration. As soon as designs other than randomized experiments are considered, the identification of average causal effects rests on the assumption of conditional unbiasedness given the considered covariates. Conditional unbiasedness holds by design in conditionally randomized experiments. However, it is not guaranteed to hold in quasi-experimental designs with self- or other-selection to treatment conditions. Unfortunately, conditional unbiasedness is a very weak causality criterion: In itself, it has no empirically testable implication. In applications, only indirect tests of conditional unbiasedness are possible by falsifying conditional unconfoundedness as a sufficient condition for conditional unbiasedness. However, tests of conditional unconfoundedness are at best indirect tests of conditional unbiasedness and have their own problems: (1) Additional distributional and functional assumptions are necessary, the hypothesized functional form of the influence of further covariates has to be specified (Steyer et al., 2009). (2) Falsification of conditional unconfoundedness does not logically imply that conditional unbiasedness does not hold, a regression can be incidentally unbiased even though it is not unconfounded. (3) Tests of conditional unconfoundedness can only falsify conditional unconfoundedness with respect to observed covariates — unconfoundedness with respect to unobserved confounders cannot be tested (see, Rosenbaum, 2002, for methods to test the sensitivity of causal inferences to unobserved potential confounders). Failure to falsify conditional unconfoundedness corroborates the notion of conditional unbiasedness with respect to the observed covariates, but does not exclude bias due to unobserved potential confounders. Nevertheless, tests for conditional unconfoundedness (and for the stronger sufficient conditions of independence and homogeneity) are currently the only available options to, at least indirectly, test the assumption of conditional unbiasedness in quasi-experimental designs. They are theoretically well understood for singlelevel models (Steyer et al., 2009); however their implementation in statistical models is still

in its developmental stage for singlelevel designs — and even more so for multilevel designs.

6.3 Research Needs

The research needs and open questions with respect to causal inference in multilevel designs and the analysis of average causal effects can be divided in two parts: (1) Obviously, further insights into the performance and adequacy of statistical implementations of the generalized ANCOVA need to be gained, especially with respect to power, the inclusion of additional covariates and alternative estimation techniques. (2) Alternative procedures need to be adapted to multilevel designs and tested for their appropriateness and sample size requirements.

6.3.1 Shortcomings of the Simulations

Both simulation studies share two obvious shortcomings that have to be addressed in future research: (1) The failure to include conditions in which the average causal effect differed from zero and (2) the inclusion of more than one covariate at the unit- and the cluster-level.

The first omission precludes comparisons of the statistical models with respect to the power of detecting average treatment effects. Power analyses and optimal design strategies are important for the planning of simulation studies and well-understood for randomized multilevel designs (e.g., Moerbeek et al., 2000; Raudenbush, 1997; Raudenbush & Liu, 2000). The results of the simulation studies, especially the relative efficiencies of the estimators, the over- and underestimation of the corresponding standard errors and the empirical type-1-error rates offer only some indications about the performance of the different implementations for the detection of treatment effects. The final word on the necessary sample sizes for adequate power requires more extensive simulation studies.

The second shortcoming is the inclusion of only one covariate at the unit-level — decomposable into its between-component Z_b and its within-component Z_w. The simulation setup was restricted on purpose to demonstrate the complexities of the implementation of the generalized ANCOVA for the most simple configuration of covariates. However, sample size requirements for more realistic setups cannot be derived from

the results of the simulation studies. In applying the generalized ANCOVA, researchers are likely to adjust for more than one covariate at the same time and have to take more complex functional forms and specifications into account. Although further research is needed to obtain additional insights for such realistic conditions, the simulations in this thesis demonstrated that the proposed methods and implementations worked in principle and showed their relative merits and deficiencies.

6.3.2 Alternatives and Extensions of the Generalized ANCOVA

Further research on causal inference in multilevel designs can build upon a sound theoretical foundation given by the general theory of causal effects. Building upon this theory, extensions of the generalized ANCOVA and the development of other adjustment procedures based on propensity scores and their adaptation to multilevel designs should be the focus of further research. Alternative statistical models that are appropriate for multilevel designs and potentially more stable than the most promising implementations of the generalized ANCOVA need to be considered: Methods that are robust to misspecifications (Kim & Frees, 2006, 2007), offer alternative ways to account for the unreliability of the cluster-means as measures for the between-component Z_b (Grilli & Rampichini, 2008) or that directly correct the standard error of the *ACE*-estimator for clustering (Hedges, 2007a) are potentially interesting options in this regard.

A second area of research is the extension of the generalized ANCOVA for multilevel designs to include latent covariates and outcome variables as well as ordered categorical variables. For singlelevel designs, the differentiation between models for latent and manifest variables has been proven to be consequential for the choice of the statistical model (Kröhne, 2009) and the same might be true for multilevel designs. The multilevel structure poses an additional challenge for latent variable models, since measurement models must be specified simultaneously for unit- and cluster-level variables increasing the sample size requirements (L. K. Muthén & Muthén, 1998-2007; Skrondal & Rabe-Hesketh, 2004). The models developed in this thesis always referred to continuous outcome variables with normally — albeit slightly heterogeneously — distributed residuals. The development of adjustment models and their statistical implementations for non-continuous outcome variables and non-normal residual distributions remain a fruitful area for further research. Other distributions of the outcome variable might not only require the choice of generalized linear models as statistical tools (Pinheiro &

Bates, 2000; Skrondal & Rabe-Hesketh, 2004), but also considerable modifications to the core concepts of the theory of causal effect in order not to lose well-defined concepts such as average and conditional causal effects. Finally, extensions of the generalized ANCOVA to non-linear intercept and effect functions are a worthwhile, but — due to the complexities involved — daunting area for further research.

As discussed before, we only focused on the generalized ANCOVA as a method for estimating the average causal effect in non-randomized multilevel designs. An extension to propensity-based adjustment methods needs to be developed for these design types and studied with respect to their efficiency and robustness to violated assumptions. While propensity-based approaches directly fit into the theoretical framework for causal inference introduced in Chapter 2 (see also Steyer et al., 2009) and the same general restrictions and assumptions are necessary to obtain valid causal inferences, propensity-based adjustment procedures and – more urgently – their statistical implementation need to be developed for multilevel designs. In order to compare the performance and robustness of propensity-based methods to the generalized ANCOVA, simulation studies and within-study comparisons (Shadish, Luellen, & Clark, 2006; Shadish, Clark, & Steiner, 2008) can be used. However, the latter approach might be difficult to implement in multilevel designs due to the comparatively high costs of adding clusters to multilevel designs (Moerbeek et al., 2000; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000).

6.4 Conclusion

When will multilevel designs be relevant in applications compared to more common singlelevel evaluation designs? Compared to singlelevel designs, multilevel designs are usually associated with higher costs due to the requirement to establish the treatment regime in more than one cluster (Moerbeek et al., 2000; Raudenbush & Liu, 2000). On the other hand, they potentially offer a higher external validity by including and averaging over different implementation characteristics at different sites (Shadish et al., 2002) As such, multilevel designs, especially designs with treatment allocation at the unit-level are often used when interventions that have been proven effective in controlled laboratory research are taken to scale and are implemented in clinical or educational practice (Shadish et al., 2002; Turpin & Sinacore, 1991). When treatments that are

naturally delivered to whole clusters are evaluated or when treatment fidelity is threatened by potential interactions between treated and untreated subjects in a cluster that cannot be modeled as cluster-covariates measurable with respect to the confounder σ-algebra $\mathfrak{C}_X$ (see also Chapter 3), a multilevel design with treatment assignment at the cluster-level offers an elegant way to evaluate treatment effects (Donner & Klar, 2000; Murray, 1998). The same is true when naturally occurring quasi-experiments or observational studies are analyzed: If one is interested in evaluating the effects of a specific teaching method as it is implemented at a certain time point in the field, the multilevel structure and the relevant cluster-covariates have to be accounted for, if a causal interpretation of the effects is desired. This was the case in the example analyses presented in Chapters 4 and 5 where the effects of naturally occurring treatments was of interest. Although these analyses were by no means complete and crucial assumptions, such as conditional unbiasedness, were not checked, they illustrated the need to account for confounding covariates with an appropriate statistical model.

One final remark remains to be made with regard to quasi-experimental designs and a common misunderstanding of discussions of causal inference for these designs (Morgan & Winship, 2007). Although we showed that causal inference in these designs is possible if the complete set of unit- and cluster-covariates that influence both treatment assignment and the outcome variable is identified and assumptions about the functional form of the treatment-specific regressions of the outcome variable on the covariates hold, this demonstration should not be misunderstood as a plea for quasi-experimental designs: Randomized (or at least conditionally randomized) experimental designs are considered as the "gold standard" (e.g., Rubin, 2008, p. 1350) for singlelevel evaluation designs and a similar point has been made with regard to multilevel designs (Donner & Klar, 2000; Murray, 1998). However, quasi-experimental designs may arise naturally when conclusions and inferences have to be made from naturally occurring treatment implementations or in situations where randomized assignment of units or clusters to treatment conditions would be unethically (e.g., randomly holding students back from advancing to the next grade level, Hong & Raudenbush, 2006). To slightly paraphrase Raudenbush (1995): Sometimes "the perfect social science study" (p. 213), a randomized experiment, is just not possible to implement and researchers have to be satisfied with quasi-experimental evidence that should not be easily discarded. It is important to study the conditions under which unbiased average causal effects can be obtained in quasi-experimental designs and the methods that can be used to do so. The derivations

and results presented in this thesis are one important step forward on this way.

A Proofs and Derivations

A.1 Equivalence of $E[g_j(Z, V, Z_b)]$ and $E[g_j(V, Z_b)]$ in Designs with Treatment Assignment at the Cluster-Level

In this section, we prove the equality

$$E[g_j(Z, V, Z_b)] = E[g_j(V, Z_b)], \tag{A.1}$$

of the expected value of the conditional effect function $E[g_j(V, Z_b)]$ from the simple adjustment model and the expected value of the conditional effect function $E[g_j(Z, V, Z_b)]$ from the full adjustment model for non-randomized multilevel designs with treatment assignment at the cluster-level as introduced in Section 5.1.

The conditional effect function $g_j(V, Z_b)$ of the *simple adjustment model* that includes only the cluster-covariate V and the between-component Z_b was defined as follows in Equation (5.4):

$$g_j(V, Z_b) = E^{\circ}_{X=j}(Y \mid V, Z_b) - E^{\circ}_{X=0}(Y \mid V, Z_b). \tag{A.2}$$

The expected value of the conditional effect function $g_j(V, Z_b)$ is equal to the difference of the expected value of the extensions of the conditional regression to Ω:

$$E[g_j(V, Z_b)] = E[E^{\circ}_{X=j}(Y \mid V, Z_b)] - E[E^{\circ}_{X=0}(Y \mid V, Z_b)]. \tag{A.3}$$

The conditional effect function $g_j(Z, V, Z_b)$ of the *full adjustment model* that includes the unit-covariate Z, as well as the cluster-covariate V and the between-component Z_b was defined as follows in Equation (5.8):

$$g_j(Z, V, Z_b) = E^{\circ}_{X=1}(Y \mid Z, V, Z_b) - E^{\circ}_{X=0}(Y \mid Z, V, Z_b). \tag{A.4}$$

The expected value of the conditional effect function $g_j(Z, V, Z_b)$ is given by the difference of the expected values of the extensions of the conditional regressions to Ω:

$$E[g_j(Z, V, Z_b)] = E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b)] - E[E^{\circ}_{X=0}(Y \mid Z, V, Z_b)]. \tag{A.5}$$

Hence, to prove Equation (A.1), we only have to prove that

$$E[E^{\circ}_{X=j}(Y \mid V, Z_b)] = E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b)], \tag{A.6}$$

for all values j of the treatment variable X. The equality in Equation (A.6) implies Equation (A.1) that we want to prove.

In order to prove Equation (A.6), we first show that *average stability* of $E(Y \mid V, Z_b)$ with respect to Z (as defined by Steyer et al., 2009) always holds in designs with treatment assignment at the cluster-level, i.e., that

$$E^{\circ}_{X=j}(Y \mid V, Z_b) = E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b) \mid V, Z_b]. \tag{A.7}$$

We will then show that Equation (A.7) implies that Equation (A.1) holds.

Average stability of $E(Y \mid V, Z_b)$ with respect to Z as defined in Equation (A.7) does not hold generally, as would

$$E_{X=j}(Y \mid V, Z_b) = E_{X=j}[E_{X=j}(Y \mid Z, V, Z_b) \mid V, Z_b]. \tag{A.8}$$

However, it will hold if

$$P(Z{=}z \mid V{=}v, Z_b{=}z_b) = P(Z{=}z \mid V{=}v, Z_b{=}z_b, X{=}j) \quad \text{for all } v \text{ and } z_b, \tag{A.9}$$

for discrete unit-covariates Z and discrete cluster-covariates V for the proof see Steyer et al. (2009). The generalization to continuous covariates is straightforward by considering the equality of the conditional distributions of the unit-covariate Z:

$$P_{Z \mid V=v, Z_b=z_b} = P_{Z \mid V=v, Z_b=z_b, X=j} \tag{A.10}$$

for all values v of V, z_b of Z_b and j of X.

In designs with treatment assignment at the cluster-level, the treatment assignment probabilities are by design independent of the unit-covariate Z given the cluster-covari-

ate V and the between-component Z_b:

$$P(X{=}j \mid Z, V, Z_b) = P(X{=}j \mid V, Z_b). \tag{A.11}$$

Equation (A.11) implies that Equation (A.9) holds without further assumptions. This can be shown as follows for discrete unit-covariates Z and discrete cluster-covariates V (the between-component Z_b is a function of the discrete cluster variable C and hence discrete by definition):

$$\begin{aligned} P(Z{=}z \mid V{=}v, Z_b{=}z_b, X{=}j) &= \frac{P(Z{=}z, V{=}v, Z_b{=}z_b, X{=}j)}{P(V{=}v, Z_b{=}z_b, X{=}j)} && \text{(A.12)} \\ &= \frac{P(X{=}j \mid Z{=}z, V{=}v, Z_b{=}z_b) \cdot P(Z{=}z, V{=}v, Z_b{=}z_b)}{P(X{=}j \mid V{=}v, Z_b{=}z_b) \cdot P(V{=}v, Z_b{=}z_b)} && \text{Factorization} \\ &= \frac{P(X{=}j \mid V{=}v, Z_b{=}z_b) \cdot P(Z{=}z, V{=}v, Z_b{=}z_b)}{P(X{=}j \mid V{=}v, Z_b{=}z_b) \cdot P(V{=}v, Z_b{=}z_b)} && \text{Eq. (A.11)} \\ &= \frac{P(Z{=}z, V{=}v, Z_b{=}z_b)}{P(V{=}v, Z_b{=}z_b).} \\ &= P(Z{=}z \mid V{=}v, Z_b{=}z_b). && \text{(A.13)} \end{aligned}$$

Equation (A.13) proves the equivalence postulated in Equation (A.9) for designs with treatment assignment at the cluster-level and implies average stability [Equation (A.7)] for discrete unit-covariates Z and cluster-covariates V. The generalization to continuous covariates using the conditional distributions of the unit-covariate Z is straightforward.

We will now use average stability to derive the equality of the two expected values of the conditional effect functions of the full and the simple adjustment model for designs with treatment assignment at the cluster-level, again assuming a discrete unit-covariate

Z and a discrete cluster-covariate V.

$$E[E^{\circ}_{X=j}(Y \mid V, Z_b)] = E[E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b) \mid V, Z_b]] \qquad \text{Eq. (A.7)} \tag{A.14}$$

$$= \sum_{v} \sum_{z_b} E[E^{\circ}_{X=j}(Y \mid Z{=}z, V{=}v, Z_b{=}z_b) \mid V{=}v, Z_b{=}z_b] \cdot P(V{=}v, Z_b{=}z_b) \tag{A.15}$$

$$= \sum_{v} \sum_{z_b} \sum_{z} E^{\circ}_{X=j}(Y \mid Z{=}z, V{=}v, Z_b{=}z_b) \cdot P(Z{=}z \mid V{=}v, Z_b{=}z_b) \cdot P(V{=}v, Z_b{=}z_b) \tag{A.16}$$

$$= \sum_{v} \sum_{z_b} \sum_{z} E^{\circ}_{X=j}(Y \mid Z{=}z, V{=}v, Z_b{=}z_b) \cdot P(Z{=}z, V{=}v, Z_b{=}z_b) \tag{A.17}$$

$$= E[E^{\circ}_{X=j}(Y \mid Z, V, Z_b)] \tag{A.18}$$

Equation (A.18) proofs the equality postulated in Equation (A.6) and directly implies the equivalence of the expected values of the conditional effect function postulated in Equation (A.1) in multilevel design with treatment assignment at the cluster-level for discrete unit-covariates Z and cluster-covariates V. Again, the generalization of the proof to continuous covariates is straightforward.

A.2 *ACE*-Estimator for the Full Adjustment Model Implemented as Multigroup Multilevel Latent Variable Model in `Mplus`

In this section, we derive the *ACE*-estimator for the implementation of the full adjustment model for multilevel designs with treatment assignment at the cluster-level as multigroup multilevel model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007). The multigroup multilevel model uses two independently estimated multilevel models for each treatment group. In order to be able to model cross-level interactions, the `TWOLEVEL RANDOM` specification in `Mplus` had to be chosen. With this option, `Mplus` models the raw scores of the unit-covariate Z instead of the within-component Z_w (in contrast to the `TWOLEVEL` specification, see also Appendix B.3). As shown in Equation (5.30) in Section 5.1.2 this complicates the resulting non-linear constraint and requires the estimation of the expected value $E(Z_b \cdot Z)$ of the product variable of the unit-covariate Z and the between-component Z_b. Due to the multigroup estimation

technique, this expected value is not directly available as a parameter in Mplus, but has to be recovered from the group-specific expected values and variances that are available as model parameters.

The multigroup multilevel model in Mplus estimates treatment group-specific regression for the control and the treatment group as given in Equations (5.27) and (5.28). We repeat these equations here, leaving out the cluster-covariate V, since it was not included in the model studied in the simulation in Section 5.2:

$$E_{X=0}(Y \mid Z, Z_b) = \gamma^*_{00} + \gamma^*_{01} Z_b + \gamma^*_{04} Z + \gamma^*_{05} Z_b \cdot Z, \tag{A.19}$$

$$\begin{aligned} E_{X=1}(Y \mid Z, Z_b) &= \gamma^*_{00} + \gamma^*_{10} + (\gamma^*_{01} + \gamma^*_{11}) Z_b \\ &\quad + (\gamma^*_{04} + \gamma^*_{14}) Z + (\gamma^*_{05} + \gamma^*_{15}) Z_b \cdot Z \end{aligned} \tag{A.20}$$

$$= \beta_0 + \beta_1 Z_b + \beta_4 Z + \beta_5 Z_b \cdot Z, \tag{A.21}$$

where the parameters β_i in Equation (A.21) are the parameters of the regression of the outcome variable Y on the unit-covariate Z and the between-component Z_b in the treatment group (X=1) as estimated in Mplus. The *ACE*-estimator is given by the expected value of the difference between Equation (A.20) and (A.19) [see Equation (5.30)] and can be expressed in terms of the treatment-group specific regression weights estimated in the multigroup multilevel latent variable model in Mplus and the expected values of the unit-covariate Z, the between-component Z_b and their product as follows:

$$\begin{aligned} ACE_{10} &= (\beta_0 - \gamma^*_{00}) + (\beta_1 - \gamma^*_{01}) \cdot E(Z) \\ &\quad + (\beta_4 - \gamma^*_{04}) \cdot E(Z) + (\beta_5 - \gamma^*_{05}) \cdot E(Z_b \cdot Z). \end{aligned} \tag{A.22}$$

This constraint has to be implemented in Mplus to obtain the *ACE*-estimator. Unfortunately, the unconditional expected values $E(Z)$ and $E(Z_b \cdot Z)$ are not directly available as parameters in Mplus and have to calculated from the treatment group-specific parameters.

The expected value $E(Z)$ of the unit-covariate Z is given by:

$$E(Z) = E(Z \mid X{=}0) \cdot P(X{=}0) + E(Z \mid X{=}1) \cdot P(X{=}1). \tag{A.23}$$

The conditional expected values $E(Z \mid X{=}j)$ in treatment group j are estimated as model parameters in Mplus. The treatment probabilities $P(X{=}j)$ have to be approximated by

the treatment-group sizes and must be manually included in the non-linear constraint.

The expected value $E(Z_b \cdot Z)$ of the product variable of the between-component Z_b and the unit-covariate Z is generally given by:

$$E(Z_b \cdot Z) = Cov(Z_b, Z) - E(Z_b) \cdot E(Z) \tag{A.24}$$

$$= Cov(Z_b, Z_b + Z_w) - E(Z) \cdot E(Z) \tag{A.25}$$

$$= Var(Z_b) - E(Z)^2. \tag{A.26}$$

The general Equation (A.24) for the expected value of product variables simplifies to Equation (A.26) by taking into account that $E(Z_b)$, the expected value of the between-component Z_b, is equal to the expected value $E(Z)$ of the unit-covariate Z [see Equation (2.8)]. Furthermore, the unit-covariate Z can be decomposed into its between-component Z_b and its within-component Z_w [see Equation (2.3)] and the covariance of Z_b and Z_w is equal to zero [see Equation (2.7)]. The squared expected value $E(Z)^2$ of the unit-covariate Z, necessary to estimate Equation (A.26) in `Mplus`, can be easily obtained from squaring Equation (A.23). The variance $Var(Z_b)$ of the between-component Z_b is generally given by:

$$Var(Z_b) = Var\big[E(Z_b \mid X)\big] + Var(\zeta), \tag{A.27}$$

where $\zeta \equiv Z_b - E(Z_b \mid X)$ is the residual of the regression $E(Z_b \mid X)$. The two components of Equation (A.27) can be expressed as model parameters in `Mplus` as follows: The variance $Var(\zeta)$ of the residual is given by

$$Var(\zeta) = Var(Z_b \mid X{=}0) \cdot P(X{=}0) + Var(Z_b \mid X{=}1) \cdot P(X{=}1), \tag{A.28}$$

where the treatment-group specific variances $Var(Z_b \mid X{=}j)$ are estimated as parameters in `Mplus` and the treatment probabilities $P(X{=}j)$ must be manually included and are approximated by the treatment-group sizes.

Finally, the variance $Var\big[E(Z_b \mid X)\big]$ of the regression of the between-component Z_b on the treatment variable X is given by:

$$Var\big[E(Z_b \mid X)\big] = \big[E(Z_b \mid X{=}0) - E(Z)\big]^2 \cdot P(X{=}0) + \big[E(Z_b \mid X{=}1) - E(Z)\big]^2 \cdot P(X{=}1), \tag{A.29}$$

where the conditional expected values $E(Z_b \mid X{=}j)$ of the between-component Z_b are estimated as parameters in `Mplus`, the unconditional expected value $E(Z)$ of the unit-covariate is given by Equation (A.23) and the treatment probabilities $P(X{=}j)$ must be entered once more manually and approximated by the treatment group sizes.

Substituting Equations (A.28) and (A.29) into Equation (A.27) and further into Equation (A.26) identifies the expected value $E(Z_b \cdot Z)$ of the product variable of the between-component Z_b and the unit-covariate Z. The `Mplus`-syntax used to implement these constraints is given in Listing B.8.

B Statistical Models

The generalized ANCOVA for conditionally randomized and quasi-experimental between-group multilevel designs developed in Chapters 4 and 5 can be implemented in different statistical models. Some of these implementations were compared with respect to their performance in finite samples in simulation studies described in Section 4.2 for designs with treatment assignment at the unit-level and in Section 5.2 for designs with treatment assignment at the cluster-level. In this appendix, we introduce the underlying statistical models in more detail and show how the average causal effect is estimated and tested in each of them. Specifically, the statistical models are (1) the generalized (singlelevel) ANCOVA as implemented in `EffectLite` (Steyer & Partchev, 2008) and the corresponding R-package `lace` (Partchev, 2007), (2) the linear mixed effect model (Laird & Ware, 1982) as implemented in the R-package `nlme` (Pinheiro & Bates, 2000; Pinheiro et al., 2008), (3) the multilevel structural equation model as implemented in `Mplus` 5.0 (Asparouhov & Muthén, 2004; L. K. Muthén & Muthén, 1998-2007), its generalization to multiple groups (B. O. Muthén, Khoo, & Gustafsson, 1997) and to latent variables for between- and within-components of manifest variables (Lüdtke et al., 2008) and (4) the two-step adjustment procedure for unbiased prediction of group-level outcomes by Croon and van Veldhoven (2007). We will briefly introduce the statistical models, discuss their merits and shortcomings, describe which adjustment models can be implemented and how this implementation is undertaken. Along the way, we will present the actual models compared in the simulation studies in Chapters 4 and 5 and give the syntax that was used to implement them.

B.1 Singlelevel Generalized ANCOVA

The generalized analysis of covariance (ANCOVA) for singlelevel analysis of average causal effects developed in Steyer et al. (2009) was used as implemented in the

R-package `lace` (Partchev, 2007). The package serves as a preprocessor for the multigroup structural equation model in `Mplus` (L. K. Muthén & Muthén, 1998-2007) and provides several tools for causal inference, including standard errors and hypotheses tests of average effects and conditional causal effects with linear effect functions (for details see Steyer & Partchev, 2008).

The implementation of the generalized ANCOVA in `lace` uses the multigroup structural equation model as implemented in either `Mplus` (L. K. Muthén & Muthén, 1998-2007) or `Lisrel` (Jöreskog & Sörbom, 1996 - 2001) to estimate and test the average causal effect. The multigroup structural equation model is given by two separate models for each treatment group j: (1) the measurement model for the vector of observed variables $\mathbf{y}_i^{(j)}$ and (2) the structural models for the vector of latent variables $\boldsymbol{\eta}_i^{(j)}$ (using `Mplus` notation, B. O. Muthén, 1998-2004, 2002):

$$\mathbf{y}_i^{(j)} = \boldsymbol{\nu}^{(j)} + \boldsymbol{\Lambda}^{(j)}\boldsymbol{\eta}_i^{(j)} + \boldsymbol{\epsilon}_i^{(j)}, \tag{B.1}$$

$$\boldsymbol{\eta}_i^{(j)} = \boldsymbol{\alpha}^{(j)} + \mathbf{B}^{(j)}\boldsymbol{\eta}_i^{(j)} + \boldsymbol{\zeta}_i^{(j)}, \tag{B.2}$$

where $\mathbf{y}_i^{(j)}$ is the p-dimensional vector of observed variables for observation i in treatment group j, $\boldsymbol{\eta}_i^{(j)}$ is the m-dimensional vector of latent variables for observation i in treatment group j, $\boldsymbol{\eta}_i^{(j)}$ is the p-dimensional vector of residual errors that is uncorrelated with the other variables, $\boldsymbol{\nu}^{(j)}$ is the p-dimensional vector of measurement intercepts in treatment group j and $\boldsymbol{\Lambda}^{(j)}$ is a $(p \times m)$-matrix of factor loadings in treatment group j. In the measurement model, $\boldsymbol{\alpha}^{(j)}$ is an m-dimensional parameter vector of intercepts in treatment group j and $\mathbf{B}^{(j)}$ is an $(m \times m)$-matrix of slopes for regressions of latent variables on latent variables in treatment group j with zero diagonal elements.

The following group-specific mean structures $\boldsymbol{\mu}_{\mathbf{y}}^{(j)}$ and covariance structures $\boldsymbol{\Sigma}_{\mathbf{y}}^{(j)}$ of the manifest variables are implied by Equations (B.1) and (B.2):

$$\boldsymbol{\mu}_{\mathbf{y}}^{(j)} = \boldsymbol{\nu}^{(j)} + \boldsymbol{\Lambda}^{(j)}(\mathbf{I} - \mathbf{B}^{(j)})^{-1}\boldsymbol{\alpha}^{(j)}, \tag{B.3}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}}^{(j)} = \boldsymbol{\Lambda}^{(j)}(\mathbf{I} - \mathbf{B}^{(j)})^{-1}\boldsymbol{\Psi}^{(j)}\boldsymbol{\Lambda}^{(j)}(\mathbf{I} - \mathbf{B}^{(j)})'^{-1}\boldsymbol{\Lambda}^{(j)\prime} + \boldsymbol{\Theta}^{(j)}, \tag{B.4}$$

where $\boldsymbol{\Psi}^{(j)}$ is the positive-definite covariance matrix of the independently, identically and multivariate normally distributed structural model residuals $\boldsymbol{\zeta}_i^{(j)}$ in treatment group j, and $\boldsymbol{\Theta}^{(j)}$ is the positive-definite covariance matrix of the independently, identically and multivariate normally distributed measurement model residuals $\boldsymbol{\epsilon}_i^{(j)}$ in treatment

group j.

The estimator of the average causal effect in the generalized ANCOVA is obtained by combining the relevant parameters of $\boldsymbol{\alpha}^{(j)}$ — including the expected values of the covariates — and $\mathbf{B}^{(j)}$ — including the regression coefficients — in a non-linear constraint (for details see, Steyer & Partchev, 2008). Standard errors for the average effect estimators are obtained with the multivariate delta method (Rao, 1973; Raykov & Marcoulides, 2004). Tests of multiple constraints are implemented using the Wald test (Wald, 1943) by imposing the constraints on the parameters of the unrestricted model. In order to account for randomly varying group sizes — in estimating the standard errors of the *ACE* and the Wald test — the variance-covariance matrix of the model parameters is augmented with the variances of the group sizes (for details see, Kröhne, 2009; Nagengast, 2006). In the simulation studies, significance tests for the *ACE* were obtained by dividing the *ACE*-estimate by its standard error and using the normal distribution as reference (see also, Nagengast, 2006). This test is equivalent to a Wald test for a single non-linear constraint (Rao, 1973).

Compared to the conventional ANCOVA implemented in the general linear model (Searle, 1971; Werner, 1997), `lace` has several advantages (see also Kröhne, 2009): (1) It is not necessary to assume homogeneity of residual variances between treatment groups by virtue of the group specific variance-covariance matrices $\boldsymbol{\Psi}^{(j)}$ and $\boldsymbol{\Theta}^{(j)}$ in multigroup structural equation models (Steyer & Partchev, 2008). (2) Correct standard errors and significance tests of the average causal effect are provided even in the presence of interactions between the treatment variable and the covariates (Kröhne, 2009). (3) Finally, the predictors are not assumed to be fixed, but are treated as stochastic and have a joint distribution with the other manifest variables in the model (Kröhne, 2009; Nagengast, 2006); their expected values are estimated as model parameters and have a joint distribution with the other model parameters. An added benefit of the use of structural equation models is the capability of modeling latent outcomes and covariates (Steyer et al., 2009; Steyer & Partchev, 2008). Since `lace` relies on the conventional structural equation model in `Mplus`, no variance components on the cluster-level can be estimated and thus all models only include fixed effects. Since the statistical background of the generalized ANCOVA in `lace` is well understood and documented (Kröhne, 2009; Steyer & Partchev, 2008), implementations of the adjustment models in `lace` serve as a standard against which more complex models can be judged, although the models implemented in `lace` were misspecified with respect to some of the

```
model.lace.z = lace(d$x, d$y, d$z,
                    control.group= "0", engine = "mplus",
                    program=mplus.executable)
```

Listing B.1: R-code for the naive adjustment model implemented in `lace`

```
model.lace = lace(d$x, d$y,
                  cbind(d$z.betw, d$z.within, d$z.betw*d$z.within),
                  control.group= "0", engine = "mplus",
                  program=mplus.executable)
```

Listing B.2: R-code for the full adjustment model implemented in `lace`

properties of the data generation procedure.

The R-package `lace` was used to implement adjustment models for both types of non-randomized multilevel designs. In case of non-randomized designs with *treatment assignment at the unit-level*, two adjustment models were implemented in `lace`:

(1) The *naive adjustment model*, that only included the unit-covariate Z and did not account for the multilevel structure of the data, was given by the following model equation

$$\mathbf{y}^{(j)} = \alpha^{(j)} + \beta^{(j)}\mathbf{z}^{(j)} + \boldsymbol{\epsilon}^{(j)}, \tag{B.5}$$

where $\mathbf{y}^{(j)}$ is the vector of observed values of the outcome variable in treatment group j, $\mathbf{z}^{(j)}$ is the vector of observed values of the unit-covariates in treatment group j, $\alpha^{(j)}$ is the intercept in treatment group j, $\beta^{(j)}$ is the regression weight of the unit-covariate in treatment group j and $\boldsymbol{\epsilon}^{(j)}$ is the residual vector in treatment group j. The syntax used to implement the naive adjustment model is given in Listing B.1.

(2) The *full adjustment model* includes as predictors: the cluster means of the unit-covariate Z, the individual values of the unit-covariate Z centered around the empirical cluster means and their product and was given by the following model equation

$$\mathbf{y}^{(j)} = \alpha^{(j)} + \beta_1^{(j)}(\mathbf{z}^{(j)} - \overline{\mathbf{z}}_c) + \beta_2^{(j)}\overline{\mathbf{z}}_c + \beta_3^{(j)}\left[(\mathbf{z}^{(j)} - \overline{\mathbf{z}}_c)\overline{\mathbf{z}}_c\right] + \boldsymbol{\epsilon}^{(j)}, \tag{B.6}$$

where $\mathbf{y}^{(j)}$ is the vector of observed values of the outcome variable in treatment group j, $\mathbf{z}^{(j)}$ is the vector of observed unit-covariates in treatment group j, $\overline{\mathbf{z}}_c$ is the vector of the observed cluster-means of the unit-covariate for each observation, $\alpha^{(j)}$ is the intercept in treatment group j, $\beta^{(j)}$ is the regression weight of the corresponding predictor in treatment group j and $\boldsymbol{\epsilon}^{(j)}$ is the residual vector in treatment group j. The syntax used to implement the full adjustment model is given in Listing B.2.

In case of non-randomized designs with *treatment assignment at the cluster-level*,

`lace` was used to estimate the same two models: (1) The *naive adjustment model* that only included the unit-covariate Z and did not take the multilevel structure of the data into account. The model equation is given in Equation (B.5), the syntax is given in Listing B.1. (2) The *full adjustment model* that used the empirical cluster means of Z, the deviations of the realized values of Z from the empirical cluster means and their product as predictors. The model equation is given in Equation (B.6), the syntax is given in Listing B.2.

B.2 Hierarchical Linear Model with Fixed Predictors

The second option for implementing the generalized ANCOVA introduced in Chapters 4 and 5 is the linear mixed effect model (Laird & Ware, 1982) that was used to implement the generalized ANCOVA as hierarchical linear regression model (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). We used the implementation of linear mixed effect models in the `R`-package `nlme` (Pinheiro et al., 2008) in all simulations, since the underlying statistical framework is well-documented (Pinheiro & Bates, 2000). All results obtained with `nlme` also apply in principle to other implementations of the hierarchical linear regression model (e.g., `lmer`, Bates et al., 2008; `MLWin`, Rasbash, Steele, Browne, & Prosser, 2005; `HLM`, Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004).

In the multilevel linear mixed effect model with one-level of nesting, $\mathbf{y}_c$, the n_c-dimensional response vector in the cth cluster is modeled as [Pinheiro & Bates, 2000, Eq. (2.1), terminology changed]

$$\mathbf{y}_c = \mathbf{W}_c\boldsymbol{\beta} + \mathbf{R}_c\mathbf{b}_c + \boldsymbol{\epsilon}_c, \qquad c = 1, \ldots, M, \tag{B.7}$$

$$\mathbf{b}_c \sim N(\mathbf{0}, \boldsymbol{\psi}), \tag{B.8}$$

$$\boldsymbol{\epsilon}_c \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{B.9}$$

The subscript c indicates that the model is replicated within each cluster c with cluster-specific vectors and matrices $\mathbf{y}_c$, $\mathbf{W}_c$, $\mathbf{R}_c$, $\mathbf{b}_c$ and $\boldsymbol{\epsilon}_c$. Where $\mathbf{W}_c$ is the $(n_c \times p)$-matrix of fixed-effects regressors, $\boldsymbol{\beta}$ is the p-dimensional vector of fixed effects, $\mathbf{R}_c$ is the $(n_c \times q)$-matrix of random-effects regressors and is usually a subset of $\mathbf{W}_c$, $\mathbf{b}_c$ is the q-dimensional vector of random effects, and $\boldsymbol{\epsilon}_c$ is the n_i-dimensional vector within-

group errors. Note, that fixed and random intercepts are modeled by including the n_c-dimensional unit-vectors $\mathbf{1}$ in either $\mathbf{W}_c$ or $\mathbf{R}_c$ respectively. Furthermore both $\mathbf{W}_c$ and $\mathbf{R}_c$ are assumed to be known matrices of fixed regressors (Pinheiro & Bates, 2000).

The random effects $\mathbf{b}_c$ are assumed to follow a multivariate-normal distribution with mean zero and a symmetric and positive-definite covariance matrix $\boldsymbol{\psi}$ [Equation (B.8)]. The residuals $\boldsymbol{\epsilon}_c$ are assumed to follow a normal distribution and to be independent of one another [Equation (B.9)]. Furthermore, independence of $\boldsymbol{b}_c$ and $\boldsymbol{\epsilon}_c$ is assumed within and between clusters. Parameter estimates and standard errors are obtained with restricted maximum likelihood (REML) estimation by the hybrid optimization method algorithm implemented in `nlme` (see Pinheiro & Bates, 2000, Section 2.2, for details).

Non-linear constraints of model parameters that identify the average causal effect can be specified and tested as a general linear hypothesis (GLH) treating the empirical mean of the unit-covariate as a fixed value. We used the implementation of the GLH in the R-package `multcomp` (Hothorn, Bretz, & Westfall, 2008). The null hypothesis of the GLH is given by (e.g., Rao, 1973; Searle, 1971; Werner, 1997):

$$H_0 : \qquad \mathbf{A}\boldsymbol{\beta} - \boldsymbol{\delta} = \mathbf{0}, \tag{B.10}$$

where $\mathbf{A}$ is a $m \times p$-matrix containing m independent linear combinations of the p model parameters, $\boldsymbol{\delta}$ contains the hypothesized values of these contrasts, and $\boldsymbol{\beta}$ is the $p \times 1$-vector of fixed regression parameters from Equation (B.7). The empirical mean $\overline{\mathbf{z}}$ of the unit-covariate Z was used as an estimator of the expected value $E(Z)$ and treated as a fixed element of the matrix $\mathbf{A}$ in the specification of the hypothesis. The normal distribution was used as reference distribution to obtain p-values for two-sided significance tests of statistical hypothesis of the average causal effect (see Hothorn et al., 2008, for a detailed description). Since all elements of the hypothesis matrix $\mathbf{A}$ are treated as fixed, the general linear hypothesis does not really test a non-linear function of the model parameters: The mean of the covariate is treated as if it were the true expected value of the covariate — not just an estimate — (Kröhne, 2009; Nagengast, 2006) and the resulting constraint is linear in the parameters $\boldsymbol{\beta}$ of the linear mixed effects model. Formally, testing the average causal effect with the general linear hypothesis is equivalent to centering the covariates around their observed mean (Aiken & West, 1996; Kröhne, 2009, for the proof).

We used `nlme` to implement the generalized ANCOVA for both types of non-ran-

```
model.lme = lme(y ~ x*z.within*z.betw,
                random = list(cluster.ID = pdDiag(~ 1 + x)),
                data = d,
                control = list(maxIter = 2000))
contrast = c(0,1,0,0,0,mean(d$z.betw),0,0)
model.lme.glh = summary(glht(model.lme,
                linfct = as.matrix(t(contrast)), alternative = "t"))
```

Listing B.3: R-code for the full adjustment model for designs with treatment assignment at the unit-level implemented in `nlme`

domized multilevel designs. For *designs with treatment assignment at the unit-level*, the full adjustment model as introduced in Section 4.1 was implemented with the empirical cluster means of the unit-covariate Z, the group-mean centered values of Z, the treatment indicator, all cross-products of these variables and a constant unit-vector as predictors in the fixed part of the model. The random part entailed the constant unit-vector and the treatment indicator. The covariance of the random effects was restricted to zero. The average causal effect was estimated and tested with the general linear hypothesis. Formally, the following model and hypothesis matrix $\mathbf{A}$ were implemented:

$$
\begin{aligned}
\mathbf{y}_c &= \mathbf{W}_c\boldsymbol{\beta} + \mathbf{R}_c\mathbf{b}_c + \boldsymbol{\epsilon}_c, \qquad \text{(B.11)}\\
\mathbf{W}_c &= \begin{pmatrix} \mathbf{1} & \mathbf{x}_c & (\mathbf{z}_c - \overline{\mathbf{z}}_c) & \overline{\mathbf{z}}_c & \mathbf{x}_c(\mathbf{z}_c - \overline{\mathbf{z}}_c) & \mathbf{x}_c\overline{\mathbf{z}}_c & (\mathbf{z}_c - \overline{\mathbf{z}}_c)\overline{\mathbf{z}}_c & \mathbf{x}_c(\mathbf{z}_c - \overline{\mathbf{z}}_c)\overline{\mathbf{z}}_c \end{pmatrix}',\\
\mathbf{R}_c &= \begin{pmatrix} \mathbf{1} & \mathbf{x}_c \end{pmatrix}',\\
\boldsymbol{\psi} &= \mathbf{D},\\
\mathbf{A} &= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \overline{\mathbf{z}} & 0 & 0 \end{pmatrix}',
\end{aligned}
$$

where $\mathbf{x}_c$ is the observed vector of the treatment indicator in cluster c, $\mathbf{z}_c$ is the vector of observed values of the unit-covariate in cluster c, $\overline{\mathbf{z}}_c$ is the empirical cluster mean of the unit-covariate in cluster c and $\overline{\mathbf{z}}$ is the observed grand mean of the unit-covariate. The covariance between the random effects was set to zero by estimating $\boldsymbol{\psi}$ as diagonal matrix $\mathbf{D}$. The syntax of the model and the specification of the corresponding GLH is given in Listing B.3.

For designs with *treatment assignment at the cluster-level*, the simple and the full adjustment model developed in Section 5.1 were implemented in `nlme`. For the *simple adjustment model*, the fixed part of the model consisted only of the empirical cluster means of the unit-covariate, the treatment indicator, the product of these two regressors and the constant. The random part consisted only of the constant. Formally, the

```
model.lme.simple = lme(y ~ x*z.betw,
                random = ~1 | cluster.ID,
                data = d,
                control = list(maxIter = 2000))
contrast = c(0,1,0,mean(d$z.betw))
model.lme.simple.glh = summary(glht(model.lme.simple,
                linfct = as.matrix(t(contrast)), alternative = "t"))
```

Listing B.4: R-code for the simple adjustment model for designs with treatment assignment at the cluster -level in `nlme`

following model and hypothesis matrix $\mathbf{A}$ were implemented:

$$\begin{aligned} \mathbf{y}_c &= \mathbf{W}_c\boldsymbol{\beta} + \mathbf{R}_c\mathbf{b}_c + \boldsymbol{\epsilon}_c, \\ \mathbf{W}_c &= \begin{pmatrix} \mathbf{1} & \mathbf{x}_c & \bar{\mathbf{z}}_c & \mathbf{x}_c\bar{\mathbf{z}}_c \end{pmatrix}', \\ \mathbf{R}_c &= \mathbf{1}, \\ \mathbf{A} &= \begin{pmatrix} 0 & 1 & 0 & \bar{\mathbf{z}} \end{pmatrix}', \end{aligned} \tag{B.12}$$

where $\mathbf{y}_c$ is the observed vector of the outcome variable in cluster c, $\mathbf{x}_c$ is the observed vector of the treatment indicator in cluster c, $\bar{\mathbf{z}}_c$ is the empirical cluster mean of the unit-covariate in cluster c and $\bar{\mathbf{z}}$ is the observed grand mean of the unit-covariate. The syntax of the model and the corresponding GLH is given in Listing B.4.

For the *full adjustment model*, the empirical cluster means of the unit-covariate Z, the group-mean centered values of Z, the treatment indicator, all cross-products of these variables and a constant unit-vector were used as predictors in the fixed part of the model. The random part of the model consisted of the constant unit-vector. Formally, the following model was estimated:

$$\begin{aligned} \mathbf{y}_c &= \mathbf{W}_c\boldsymbol{\beta} + \mathbf{R}_c\mathbf{b}_c + \boldsymbol{\epsilon}_c, \\ \mathbf{W}_c &= \begin{pmatrix} \mathbf{1} & \mathbf{x}_c & (\mathbf{z}_c - \bar{\mathbf{z}}_c) & \bar{\mathbf{z}}_c & \mathbf{x}_c(\mathbf{z}_c - \bar{\mathbf{z}}_c) & \mathbf{x}_c\bar{\mathbf{z}}_c & (\mathbf{z}_c - \bar{\mathbf{z}}_c)\bar{\mathbf{z}}_c & \mathbf{x}_c(\mathbf{z}_c - \bar{\mathbf{z}}_c)\bar{\mathbf{z}}_c \end{pmatrix}', \\ \mathbf{R}_c &= \mathbf{1}, \\ \mathbf{A} &= \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \bar{\mathbf{z}} & 0 & 0 \end{pmatrix}', \end{aligned} \tag{B.13}$$

where $\mathbf{y}_c$ is the observed vector of the outcome variable in cluster c, $\mathbf{x}_c$ is the observed vector of the treatment indicator in cluster c, $\mathbf{z}_c$ is the vector of observed values of the unit-covariate in cluster c, $\bar{\mathbf{z}}_c$ is the empirical cluster mean of the unit-covariate in cluster c and $\bar{\mathbf{z}}$ is the observed grand mean of the unit-covariate. The syntax of the model and

```
model.lme = lme(y ~ x*z.within*z.betw,
                random = ~ 1 | cluster.ID,
                data = d,
                control = list(maxIter = 2000))
contrast = c(0,1,0,0,0,mean(d$z.betw),0,0)
model.lme.glh = summary(glht(model.lme,
                linfct = as.matrix(t(contrast)), alternative = "t"))
```

Listing B.5: R-code for the full adjustment model for designs with treatment assignment at the cluster-level in `nlme`

the corresponding GLH is given in Listing B.5.

B.3 Multilevel Structural Equation Models

While the implementation of the generalized ANCOVA in a conventional linear mixed effects model assumes fixed regressors $\mathbf{W}_c$ and $\mathbf{R}_c$, multilevel structural equation models allow the specification of models with stochastic predictors. The implementation of the hierarchical structural equation model in `Mplus` 5.0 (L. K. Muthén & Muthén, 1998-2007) allows stochastic predictors and gives estimates for the expected values and variance-covariance matrix of the predictors. The second advantage of the multilevel linear structural equation model is its capacity to model between- and within-cluster components of level-1-variables as latent variables (Lüdtke et al., 2008; L. K. Muthén & Muthén, 1998-2007).

The general hierarchical structural equation model in `Mplus` is given by L. K. Muthén and Muthén (1998-2007), its extension to models with random slopes and the implementation of the FIML algorithm used to estimate the models by Asparouhov and Muthén (2004) and B. O. Muthén and Asparouhov (2008). The extension for multigroup multilevel models follows the presentation in B. O. Muthén et al. (1997).

The hierarchical structural equation model in `Mplus` separates the vector of observed variables $\mathbf{v}_{ci}$ for individual i in cluster c in the cluster-level variables $\mathbf{z}_c$ and individual-specific variables $\mathbf{y}_{ci}$ and $\mathbf{x}_{ci}$ (B. O. Muthén, 1998-2004, p. 41):

$$\mathbf{v}_{ci} = \begin{pmatrix} \mathbf{z}_c \\ \mathbf{y}_{ci} \\ \mathbf{x}_{ci} \end{pmatrix} = \mathbf{v}_c^* + \mathbf{v}_{ci}^* = \begin{pmatrix} \mathbf{z}_c \\ \mathbf{v}_{yc}^* \\ \mathbf{v}_{xc}^* \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_{yci}^* \\ \mathbf{v}_{xci}^* \end{pmatrix}, \tag{B.14}$$

where $\mathbf{y}_{ci}$ is the vector of manifest indicators of latent variables and $\mathbf{x}_{ci}$ is the vector

of manifest covariates not included in measurement models for latent variables. The vector $\mathbf{v}_c^*$ contains the between-cluster components of the variables and the vector $\mathbf{v}_{ci}^*$ contains the within-cluster components of the corresponding variables on the unit-level. Both vectors are unobservable, latent variables. `Mplus` offers two ways of modeling the elements of these vectors – either using group-mean centering of unit-level variables and the corresponding group means as additional predictors (Enders & Tofighi, 2007; Kreft et al., 1995; Raudenbush & Bryk, 2002) or using a full-information latent variable approach (Lüdtke et al., 2008).

The multilevel structural equation model can be best characterized by considering the within- and the between-cluster model separately (L. K. Muthén & Muthén, 1998-2007). The within-cluster measurement model and the within-cluster structural model are given by B. O. Muthén [1998-2004, Eq. (192) and (193)]. The subscript c on $\mathbf{\Lambda}_{Wc}$ and $\mathbf{B}_{Wc}$ indicates that the coefficient matrices can vary between clusters to allow for random coefficients that vary between clusters in both the measurement and the structural model (see B. O. Muthén & Asparouhov, 2008):

$$\begin{pmatrix} \mathbf{v}_{yci}^* \\ \mathbf{v}_{xci}^* \end{pmatrix} = \mathbf{\Lambda}_{Wc}\boldsymbol{\eta}_{Wci} + \boldsymbol{\epsilon}_{Wci}, \tag{B.15}$$

$$\boldsymbol{\eta}_{Wci} = \mathbf{B}_{Wc}\boldsymbol{\eta}_{Wci} + \boldsymbol{\zeta}_{Wci}. \tag{B.16}$$

The within-model is completed by the two positive-definite covariance matrices $\mathbf{\Psi}_W$ and $\mathbf{\Theta}_W$ of the identically, independently and multivariate normally distributed within-cluster residuals $\boldsymbol{\zeta}_{Wci}$ and $\boldsymbol{\epsilon}_{Wci}$ that are assumed to be constant across clusters:

$$Var(\boldsymbol{\zeta}_{Wci}) = \mathbf{\Psi}_W, \tag{B.17}$$

$$Var(\boldsymbol{\epsilon}_{Wci}) = \mathbf{\Theta}_W. \tag{B.18}$$

The expected value of the within-cluster latent variable vector $E(\boldsymbol{\eta}_{Wci})$ is by default equal to $\mathbf{0}$ (B. O. Muthén et al., 1997).

The between-cluster measurement model and the between-cluster structural model are given by B. O. Muthén [1998-2004, Eq. (190) and (191)]:

$$\mathbf{v}_c^* = \boldsymbol{\nu}_B + \mathbf{\Lambda}_B\boldsymbol{\eta}_{B_c} + \boldsymbol{\epsilon}_{Bc}, \tag{B.19}$$

$$\boldsymbol{\eta}_{Bc} = \boldsymbol{\alpha}_B + \mathbf{B}_B\boldsymbol{\eta}_{Bc} + \boldsymbol{\zeta}_{Bc}. \tag{B.20}$$

When $\mathbf{\Lambda}_{Wc}$ or $\mathbf{B}_{Wc}$ are random coefficients that vary between clusters, they are included as elements of the latent variable vector $\boldsymbol{\eta}_{Bc}$ and modeled as such on the cluster-level (Asparouhov & Muthén, 2004). Thus, Equation (B.20) can be used to specify cross-level interactions. The between-model is completed by the expected value vector $\boldsymbol{\alpha}$ of the latent between-cluster variables $\boldsymbol{\eta}_{Bc}$ and the positive semi-definite variance-covariance matrices $\mathbf{\Psi}_B$ and $\mathbf{\Theta}_B$ of the identically, independently and multivariate normally distributed between-cluster residuals $\boldsymbol{\zeta}_{Bc}$ and $\boldsymbol{\epsilon}_{Bc}$:

$$E(\boldsymbol{\eta}_{Bc}) = \boldsymbol{\alpha}, \tag{B.21}$$

$$Var(\boldsymbol{\zeta}_{Bc}) = \mathbf{\Psi}_B, \tag{B.22}$$

$$Var(\boldsymbol{\epsilon}_{Bc}) = \mathbf{\Theta}_B. \tag{B.23}$$

The implied variance and covariance structure of the model cannot be expressed in closed form, if random slopes at the within-level are included in the model (Asparouhov & Muthén, 2004; B. O. Muthén, 2002). The model is estimated with a full-information maximum likelihood (FIML) estimator using an accelerated EM-algorithm (Asparouhov & Muthén, 2004; B. O. Muthén & Asparouhov, 2008). The `Mplus` implementation of the hierarchical linear structural equation model offers two options for estimating the variance-covariance matrices of the model parameters (L. K. Muthén & Muthén, 1998-2007): the standard inverse of the Hessian (see, Eliason, 1993; Rao, 1973) and a sandwich-type estimator (Huber, 1967; White, 1982) that is robust to violations of the multivariate normal distribution of the residuals. In the simulation, only the latter (and default) estimator of the variance-covariance matrix of the model parameters was used. Exploratory simulations indicated that the standard estimator yielded similar results in terms of convergence and relative bias.

The hierarchical structural equation model in `Mplus` accounts for cross-level interactions by including random slopes in the latent variable vector $\boldsymbol{\eta}_{Bc}$. In order to model interactions between discrete manifest variables (e.g., the treatment variable X) at the cluster-level and other continuous (latent) variables, a multigroup multilevel structural equation model is used. This is done by replicating Equations (B.15) to (B.23) for each subgroup j of the discrete manifest variable, thereby estimating a separate model for each treatment group with the appropriate equality constraints in place (B. O. Muthén et al., 1997). Thus the effect of cluster-covariates as well as residual variances and covariances on both the unit- and the cluster-level can vary between the treatment groups

allowing heteroscedastic error structures. However, this method is restricted to situations in which each cluster can be clearly assigned to one of the values of the discrete variable (L. K. Muthén & Muthén, 1998-2007). This restricts the applicability of the multigroup multilevel model to designs with treatment assignment at the cluster-level.

`Mplus` offers two options for estimating and testing non-linear constraints of model parameters: (1) The constrained model can be estimated directly and compared to the unrestricted model using a likelihood ratio test. (2) Alternatively, the constraint and its standard error can be computed from the unconstrained model using the multivariate delta-method (Rao, 1973; Raykov & Marcoulides, 2004). Only the latter method was used to obtain *ACE*-estimates and their standard errors for the generalized ANCOVA introduced in Chapters 4 and 5: Standard errors for the *ACE*-estimator were obtained with the multivariate delta-method as implemented in `Mplus`, p-values and significance tests were obtained by using the standard normal distribution as reference distribution.

The `Mplus`-implementation of the hierarchical structural equation model is advantageous for the specification of adjustment models for multilevel designs with non-randomized treatment allocation in two regards, even for models that do not include latent variables: (1) The expected values of covariates are estimated as model parameters in the vector $\boldsymbol{\alpha}$ and have a joint distribution with other model parameters. They can be used to define non-linear constraints of the parameters of the adjustment model. Covariates and other predictors are treated as stochastic — in contrast to the conventional hierarchical linear model, that treats predictors as fixed — appropriate for quasi-experimental designs with self- or other-selection and other designs in which the realized values of the predictors vary from sample to sample (Nagengast, 2006; Sampson, 1974; Gatsonis & Sampson, 1989). (2) The option to model within- and between-cluster components of manifest variables as latent variables is an additional advantage of the `Mplus` model. This implementation leads to unbiased estimators of the between-effect of a variable, while the commonly used method of group-mean centering with the observed group-means leads to biased estimates of this effect [Lüdtke et al., 2008, see also Equation (2.14)]. (3) Finally, by means of the multigroup multilevel structural equation model, all parameters can vary between treatment groups, thereby allowing for variance heterogeneity and interactions between the treatment variable and covariates at the cluster-level — even if within- and between-components of the manifest variables are modeled as latent variables. As mentioned before, the latter advantage requires that every cluster is uniquely assigned to one treatment group only: It

is thus only suitable for designs with treatment allocation at the cluster-level. As evident from the simulation studies, the `Mplus`-model implementation was rather instable in estimating the adjustment models. Exploratory simulations showed that stability and convergence could be increased by freeing random slopes that were not significantly different from zero in the simulations, whenever possible in `Mplus` 5.0 (see also Asparouhov & Muthén, 2004), the current version at the time of the simulation studies. The corresponding *ACE*-estimators were not influenced by this decision and, hence, the corresponding models were used in the simulation studies. The most recent version of `Mplus` 5.2 includes a significantly improved and presumably more stable optimization routine for the multigroup multilevel latent variable model that should no longer exhibit these problems (B.O. Muthen, personal communication, November 29, 2008).

The generalized ANCOVAs for non-randomized multilevel designs developed in Chapters 4 and 5 were implemented in the following ways to take advantage of the unique capabilities of the `Mplus` modeling framework. For *designs with treatment assignment at the unit-level*, the *full adjustment model* could only be specified as a singlegroup model. Thus, it was not possible to take advantage of the multilevel latent variable framework and model the interaction between the treatment variable X and the latent between- and within-components. Consequently, the adjustment model had to be specified using cluster-mean centered values of the unit-covariate Z and the corresponding empirical cluster means as predictors in order to, at least, account for the stochasticity of the unit-covariate. Specifically, the full adjustment model was specified by using the intercept, the cluster-mean centered values of Z, the treatment indicator and their product as predictors on the unit-level. The cluster-means of the unit-covariate Z were used as predictors on the cluster-level. Cross-level interactions were allowed between all variables and residual random intercepts and slopes were included. The cluster-level

```
ANALYSIS:    TYPE = TWOLEVEL RANDOM;
             ESTIMATOR = MLR;
MODEL:       %Within%
             beta1 | y on Ix1;
             beta2 | y on z_with;
             beta3 | y on zwx;
             %Between%
             [y](gamma00);
             [beta1](gamma10);
             [beta2](gamma20);
             [beta3](gamma30);
             y on z_betw (gamma01);
             beta1 on z_betw (gamma11);
             beta2 on z_betw (gamma21);
             beta3 on z_betw (gamma31);
             z_betw;
             [z_betw] (exp_z);
MODEL CONSTRAINT:
             new(ACE);
             ACE =  gamma10 + gamma11*exp_z;
```

Listing B.6: `Mplus`-syntax for the full adjustment model for designs with treatment assignment at the unit-level.

residuals were not allowed to correlate. The following formal model was implemented:

$$\mathbf{y}_c - \overline{\mathbf{y}}_c = \mathbf{B}_{Wc}\boldsymbol{\eta}_{Wc} + \boldsymbol{\epsilon}_{Wc}, \tag{B.24}$$

$$\boldsymbol{\eta}_{Wc} = \begin{pmatrix} \mathbf{x}_c & (\mathbf{z}_c - \overline{\mathbf{z}}_c) & \mathbf{x}_c(\mathbf{z}_c - \overline{\mathbf{z}}_c) \end{pmatrix}', \tag{B.25}$$

$$\overline{\mathbf{y}}_c = \boldsymbol{\alpha}_B + \mathbf{B}_B\boldsymbol{\eta}_{Bc} + \boldsymbol{\zeta}_{Bc}, \tag{B.26}$$

$$\boldsymbol{\eta}_{Bc} = \begin{pmatrix} \overline{\mathbf{y}}_c & \mathbf{B}_{Wc} & \overline{\mathbf{z}}_c \end{pmatrix}', \tag{B.27}$$

$$\boldsymbol{\zeta}_{Bc} = \begin{pmatrix} \zeta_{\mathbf{y}_c} & \zeta_{\mathbf{x}_c} & \zeta_{\mathbf{z}_c} & \zeta_{\mathbf{x}_c\mathbf{z}_c} & 0 \end{pmatrix}, \tag{B.28}$$

$$\boldsymbol{\Psi}_B = \boldsymbol{D}, \tag{B.29}$$

where $\mathbf{y}_c$ is the observed vector of the outcome variable in cluster c, $\overline{\mathbf{y}}_c$ is the empirical cluster mean of the outcome variable in cluster c, $\mathbf{x}_c$ is the observed vector of the treatment indicator in cluster c, $\mathbf{z}_c$ is the vector of observed values of the unit-covariate in cluster c and $\overline{\mathbf{z}}_c$ is the empirical cluster mean of the unit-covariate in cluster c. The full model syntax, including the non-linear constraint to estimate the *ACE* is given in Listing B.6.

The `Mplus`-framework offered significantly more options for *designs with treatment assignment at the cluster-level*. The following adjustment models were specified: In order to take advantage of the capability to model stochastic covariates, a *singlegroup multilevel model* with the cluster-mean centered unit-covariate Z as predictor at the unit-

```
ANALYSIS:    TYPE = TWOLEVEL RANDOM;
             ESTIMATOR = MLR;
MODEL:       %Within%
             beta1 | y on z_within;
             %Between%
             y on z_betw (gamma_02);
             y on Ix1  (gamma_01);
             y on xz (gamma_04);
             beta1 on Ix1 (gamma_11);
             beta1 on z_betw (gamma_12);
             beta1 on xz  (gamma_14);
             [y];
             [z_betw](E_betw_Z);
             [Ix1]; [xz]; [beta1];
             Ix1 with z_betw;
             Ix1 with xz;
             xz with z_betw;
MODEL CONSTRAINT:
             new(ACE);
             ACE = gamma_01 + gamma_04 * E_betw_Z;
```

Listing B.7: `Mplus`-syntax for the full adjustment model for designs with treatment assignment at the cluster-level as singlegroup multilevel model

level, the empirical cluster means of the unit-covariate Z as predictor and the treatment indicator variable and their cross-product at the cluster-level and the cross-level interactions as fixed effects. Residual variance component was estimated for the intercept and the unit-covariate. The correlation between the cluster-level residuals was restricted to zero. Formally, the following model was implemented:

$$\mathbf{y}_c - \overline{\mathbf{y}}_c = \mathbf{B}_{Wc}\boldsymbol{\eta}_{Wc} + \boldsymbol{\epsilon}_{Wc}, \tag{B.30}$$

$$\boldsymbol{\eta}_{Wc} = (\mathbf{z}_c - \overline{\mathbf{z}}_c), \tag{B.31}$$

$$\overline{\mathbf{y}}_c = \boldsymbol{\alpha}_B + \mathbf{B}_B\boldsymbol{\eta}_{Bc} + \boldsymbol{\zeta}_{Bc}, \tag{B.32}$$

$$\boldsymbol{\eta}_{Bc} = \begin{pmatrix} \overline{\mathbf{y}}_c & \mathbf{B}_{Wc} & \mathbf{x}_c & \overline{\mathbf{z}}_c & \mathbf{x}_c\overline{\mathbf{z}}_c \end{pmatrix}', \tag{B.33}$$

$$\boldsymbol{\zeta}_{Bc} = \begin{pmatrix} \boldsymbol{\zeta}_{\mathbf{y}_c} & \boldsymbol{\zeta}_{\mathbf{z}_c} & 0 & 0 & 0 \end{pmatrix}, \tag{B.34}$$

$$\boldsymbol{\Psi}_B = \boldsymbol{D}, \tag{B.35}$$

where $\mathbf{y}_c$ is the observed vector of the outcome variable in cluster c, $\overline{\mathbf{y}}_c$ is the empirical cluster mean of the outcome variable in cluster$c$, $\mathbf{x}_c$ is the observed vector of the treatment indicator in cluster c, $\mathbf{z}_c$ is the vector of observed values of the unit-covariate in cluster c and $\overline{\mathbf{z}}_c$ is the empirical cluster mean of the unit-covariate in cluster c. The full model syntax, including the non-linear constraint to estimate the *ACE* is given in Listing B.7.

In order to profit from the latent variable modeling capacities for the within- and between-components of the unit-covariate Z, a second implementation of the full adjustment model used the *multigroup multilevel latent variable model*: For each treatment group, a multilevel model with the within-component of the unit-covariate Z as predictor at the unit-level and the between-component as predictor at the cluster-level as fixed effects and uncorrelated variance components for the intercept and the effect of the within-component were specified. Additionally, this model included estimates of the group-specific expected values of the between-component of the unit-covariate Z.

$$\mathbf{v}_{yci}^{*(j)} = \mathbf{B}_{Wc}^{(j)}\boldsymbol{\eta}_{Wc}^{(j)} + \boldsymbol{\epsilon}_{Wc}^{(j)}, \tag{B.36}$$

$$\boldsymbol{\eta}_{Wc}^{(j)} = \mathbf{v}_{zci}^{*(j)}, \tag{B.37}$$

$$\mathbf{v}_{c}^{*(j)} = \boldsymbol{\alpha}_{B}^{(j)} + \mathbf{B}_{B}^{(j)}\boldsymbol{\eta}_{Bc}^{(j)} + \boldsymbol{\zeta}_{Bc}^{(j)} \tag{B.38}$$

$$\boldsymbol{\eta}_{Bc}^{(j)} = \begin{pmatrix} \mathbf{v}_{yc}^{*(j)} & \mathbf{B}_{Wc}^{(j)} & \mathbf{v}_{zc}^{*(j)} \end{pmatrix}', \tag{B.39}$$

$$\boldsymbol{\zeta}_{Bc}^{(j)} = \begin{pmatrix} \boldsymbol{\zeta}_{\mathbf{v}_{yc}^{*(j)}}^{(j)} & \boldsymbol{\zeta}_{\mathbf{v}_{zci}^{*(j)}}^{(j)} & 0 \end{pmatrix}, \tag{B.40}$$

$$\boldsymbol{\Psi}_{B}^{j} = \boldsymbol{D}, \tag{B.41}$$

where $\mathbf{v}_{yci}^{*(j)}$ is the vector of within-components of the outcome variable in treatment group j and cluster c, $\mathbf{v}_{yc}^{*(j)}$ is the vector of between-components of the outcome variable in treatment group j and cluster c, $\mathbf{v}_{zci}^{*(j)}$ is the vector of within-components of the unit-covariate in treatment group j and cluster c and $\mathbf{v}_{zc}^{*(j)}$ is the vector of between-components of the unit-covariate in treatment group j and cluster c. The full model syntax, including the non-linear constraint to estimate the *ACE* is given in Listing B.8. The derivation of the non-linear constraint for the *ACE*-estimator and the calculation for the expected value $E(Z \cdot Z_b)$ has been given in Appendix A.2.

B.4 Adjustment Procedure of Croon and van Veldhoven (2007)

Croon and van Veldhoven (2007) developed a two-step adjustment procedure to estimate the linear regression of a cluster-level outcome variable on the cluster-means of a unit-level predictor without bias. Their method is based on first adjusting the empirical cluster means of the unit-covariate Z using basic ANOVA formulas. The so-adjusted

```
ANALYSIS:   TYPE = TWOLEVEL RANDOM;
            ESTIMATOR = MLR;
MODEL:      %Within%
            beta1 | y on z;
            %Between%
            y on z (gamma01);
            beta1 on z (gamma03);
            [z] (Ez0);
            [y] (gamma00);
            [beta1] (gamma02);
MODEL treat:
            %Within%
            %Between%
            y on z (gamma11);
            beta1 on z (gamma13);
            [z] (Ez1);
            [y] (gamma10);
            [beta1] (gamma12);
MODEL CONSTRAINT:
            new(EZ);
            new(Vareps);
            new(VarZbX);
            new(VarZb);
            new(EZbz);
            new(ACE);
            EZ = (SIZE0 * Ez0) + (SIZE1 * Ez1);
            Vareps = (SIZE0 * Varz0) + (SIZE1 * Varz1);
            VarZbX = (SIZE0 * ((Ez0 - EZ)^2))
                   + (SIZE1 * ((Ez1 - EZ)^2));
            VarZb = VarZbX + Vareps;
            EZbz = VarZb + EZ^2;
            ACE = (gamma10 - gamma00)
                + (gamma11 - gamma01) * EZ
                + (gamma12 - gamma02) * EZ
                + (gamma13 - gamma03) * EZbz;
```

Listing B.8: `Mplus`-syntax for the full adjustment model for designs with treatment-allocation at the cluster-level as multigroup multilevel latent variable model. `SIZE0` and `SIZE1` are the relative treatment group sizes that were adapted for each data set

cluster means are then used to predict cluster-level outcome variables and to obtain an unbiased and consistent estimator of the between-cluster effect of a unit-level predictor with a simple linear model.

Their procedure consists of four steps: (1) Two weight matrices $\mathbf{W}_{g1}$ and $\mathbf{W}_{g2}$ are estimated using unbiased (or at least) consistent estimators of their components, mean vectors and covariance matrices of the between- and within-components (see Croon & van Veldhoven, 2007, for details). The cluster-specific weight matrices $\mathbf{W}_{g1}$ and $\mathbf{W}_{g2}$ are given by the following equations:

$$\mathbf{W}_{g1} = (\boldsymbol{\Sigma}_{\xi\xi} + \boldsymbol{\Sigma}_{\nu\nu}/n_g - \boldsymbol{\Sigma}_{\xi z}\boldsymbol{\Sigma}_{\xi\xi}^{-1}\boldsymbol{\Sigma}_{z\xi})^{-1}(\boldsymbol{\Sigma}_{\xi\xi} - \boldsymbol{\Sigma}_{\xi z}\boldsymbol{\Sigma}_{\xi\xi}^{-1}\boldsymbol{\Sigma}_{z\xi}), \tag{B.42}$$

$$\mathbf{W}_{g2} = \boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{z\xi}(\mathbf{I} - \mathbf{W}_{g1}), \tag{B.43}$$

where $\mathbf{\Sigma}_{\xi\xi}$ is the variance-covariance matrix of the of the between-components of the unit-level variables, $\mathbf{\Sigma}_{\nu\nu}$ is the variance-covariance matrix of the of the within-components of the unit-level variable, $\mathbf{\Sigma}_{zz}$ is the variance-covariance matrix of the cluster-level variables and $\mathbf{\Sigma}_{z\xi}$ is the covariance matrix of the between-components of cluster-level variables and between-components of unit-level variables. If cluster sizes differ, estimators for $\mathbf{W}_{g1}$ and $\mathbf{W}_{g2}$ must be obtained separately for each cluster.

(2) The weight matrices are then used to calculate the adjusted cluster mean of the unit-level variables $\tilde{\mathbf{x}}_g$ in cluster g for all clusters according to the following formula:

$$\tilde{\mathbf{x}}'_g = \boldsymbol{\mu}'_{\xi}(\mathbf{I} - \mathbf{W}_{g1}) + \bar{\mathbf{x}}'_g\mathbf{W}_{g1} + (\mathbf{z}_g - \boldsymbol{\mu}_z)'\mathbf{W}_{g2}. \tag{B.44}$$

(3) A linear model for the regression of the cluster-level outcome variable $\mathbf{y}_g$ on the adjusted group means of the unit-level variables $\tilde{\mathbf{x}}$ and the cluster-level variables $\mathbf{z}_g$ is specified and estimated with an OLS regression. The regression is given by:

$$\mathbf{y}_g = \beta_0 + \tilde{\mathbf{x}}'\boldsymbol{\beta}_1 + \mathbf{z}'_g\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_g. \tag{B.45}$$

(4) The fourth and final step consists of the calculation of heteroscedasticity-consistent estimators of the standard errors after (White, 1980). Heteroscedasticity of the error variable $\boldsymbol{\epsilon}_g$ arises when clusters differ in size.

Croon and van Veldhoven (2007) confined themselves to modeling true cluster-level variables as dependent variables, whereas the adjustment models introduced in Chapter 5 were developed for the outcome variable Y measured at the unit-level. Therefore their adjustment procedure had to be slightly modified to implement the adjustment models. Furthermore, Croon and van Veldhoven did not explicitly include interactions of the predictors into their derivation; such interactions between the treatment variable X and the adjusted cluster means of Z, however, had to be included as predictors for the generalized ANCOVA. Furthermore, the adjustment procedure relies on the general linear model and thus implicitly assumes fixed predictors. Since it only includes variables at the cluster-level, it is not a fully efficient procedure (Lüdtke et al., 2008). On the positive side, it requires only some easy calculations, is considerably fast and has yielded promising results in simulation studies (Croon & van Veldhoven, 2007; Lüdtke et al., 2008).

The adjustment procedure by Croon and van Veldhoven (2007) could only be used to

```
model.croon = lm(y.tilde ~ x*z.tilde)
contrast.croon = c(0,1,0,mean(d$z.betw))
glht(model.croon, linfct =
                    as.matrix(t(contrast.croon)), alternative = "t")
```

Listing B.9: R-syntax for the adjustment procedure of Croon and van Veldhoven (2007) using the general linear hypothesis

```
model.lace.croon = lace(x, y.tilde, z.tilde,
                        control.group= "0", engine = "mplus",
                        program=mplus.executable)
```

Listing B.10: R-syntax for the adjustment procedure of Croon and van Veldhoven (2007) using `lace`

implement the simple adjustment model for *designs with treatment assignment at the cluster-level* as introduced in Section 5.1. It was used with the following modifications: In order to not only correct the cluster means of the unit covariate Z, but also of the outcome variable Y for unreliability, Step 1 and 2 were applied to both the vector of the outcome variable $\mathbf{y}$ and the vector of the unit-covariate $\mathbf{z}$. The vector of adjusted cluster means of the outcome variable $\tilde{\mathbf{y}}$ was then regressed on the vector of the adjusted cluster means of the unit-covariate $\tilde{\mathbf{z}}$, the treatment variable $\mathbf{x}$ and the product of $\tilde{\mathbf{z}}$ and $\mathbf{x}$ using the general linear model. The *ACE*-estimator, its standard error and the significance test of the null hypothesis were then obtained with the general linear hypothesis using the mean of the adjusted cluster means of the unit-covariate Z in the specification of the hypothesis matrix $\mathbf{A}$. Step 4, the calculation of a heteroscedasticity-consistent covariance matrix estimator for unequal cluster sizes, was omitted since cluster sizes differed only moderately in the present simulations and Croon and van Veldhoven (2007) reported that there would be not much efficiency gained by using the heteroscedasticity-consistent estimator of the standard error. Specifically, the following general linear model was estimated:

$$\tilde{\mathbf{y}}_g = \beta_0 + \tilde{\mathbf{z}}_g'\beta_1 + \mathbf{x}_\mathbf{g}'\beta_2 + (\tilde{\mathbf{z}}_g\mathbf{x}_\mathbf{g})'\beta_3 + \boldsymbol{\epsilon}_g. \tag{B.46}$$

The model syntax for the general linear model and the GLH as implemented in the R-package `multcomp` (Hothorn et al., 2008) with the normal distribution as reference distribution are given in Listing B.9. The complete R-syntax for the adjustment procedure of Croon and van Veldhoven (2007) is given in Listing B.11.

In order to account for the stochasticity of the unit-covariate, the adjusted cluster means $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{z}}$ were also used to obtain estimates, standard errors and significance tests

for the average causal effect from `lace` (Partchev, 2007). Treatment group specific regressions of adjusted cluster-means of the outcome variable $\tilde{\mathbf{y}}$ on the adjusted cluster means of the unit-covariate $\tilde{\mathbf{z}}$ were estimated. The *ACE*-estimator was obtained as a non-linear constraint of the model parameters; its standard error by means of the multivariate delta method. Specifically, the following model was estimated:

$$\tilde{\mathbf{y}}_g^{(j)} = \boldsymbol{\alpha}^{(j)} + \boldsymbol{\beta}^{(j)}\tilde{\mathbf{z}}_g + \boldsymbol{\epsilon}^{(j)}. \tag{B.47}$$

The model syntax is given in Listing B.10.

```
croon.adjust <- function(data1, data2)
{
     ng <- dim(data1)[1]
     nt <- dim(data2)[1]
     g <- data2[,1]
     ns <- tapply(rep(1,nt),g,sum)
     mgroup <- dim(data1)[2]-1
     mind <- dim(data2)[2]-1
     x <- as.matrix(data2[,1+(1:mind)])
     z <- as.matrix(data1[,1:mgroup])
     y <- data1[,1+mgroup]
     mux <- apply(x,2,mean)
     muz <- apply(z,2,mean)
     dz <- z - matrix(rep(muz,ng),ncol=mgroup,byrow=T)
     xmean <- matrix(0,ng,mind)
     for (i in 1:mind){
          xmean[,i]<- tapply(x[,i],g,mean)
      }
      vv <- var(cbind(z,xmean))
      ind1 <- 1:mgroup
      ind2 <- mgroup + (1:mind)
      vzz <- vv[ind1,ind1,drop=F]
      vzxi <- vv[ind1,ind2,drop=F]
      mm <- matrix(rep(c(xmean),rep(ns,mind)),ncol=mind)
      d <- x -mm
      mse <- t(d) %*% d/(nt-ng)
      vu <- mse
      d <- mm - matrix(rep(mux,nt),ncol=mind,byrow=T)
      msa <- t(d) %*% d /(ng-1)
      cc <- (nt - sum(ns^2)/nt)/(ng-1)
      vxi <- (msa-mse)/cc
      xtilde <- matrix(0,ng,mind)
      r2 <- solve(vzz,vzxi)
      r1 <- vxi - t(vzxi) %*% r2
      id <- diag(mind)
      for (i in 1:ng){
           p <- solve(r1 +vu/ns[i],r1)
           q <- r2 %*% (id-p)
           xtilde[i,] <- xmean[i,] %*% p +mux %*% (id-p) +dz[i,] %*% q
      }
      daf <- data.frame(z,xmean,xtilde,y)
      dimnames(daf)[[2]]<- c("ID", "y.mean", "z.mean",
                             "y.tilde", "z.tilde", "x")
      daf$x.z.tilde = daf$x * daf$z.tilde
      u <- cbind(rep(1,ng),as.matrix(cbind(daf[,6],daf[,5], daf[,7])))
      e <- res$residuals
      p <- solve(t(u) %*% u)
      h <- diag(u %*% p %*% t(u))
      d <- e^2/(1-h)
      v <- p %*% t(u) %*% diag(d) %*% u %*% p
      return(list(
            adjusted.cov = v ,
            daf = daf)
            )
      }
```

Listing B.11: R-function to obtain the reliability corrected cluster means of the outcome and the unit-covariate of Croon and van Veldhoven (2007)

C Data Generation Procedures

In this appendix, the implementation of the data generation procedures in R (R Development Core Team, 2008) and the actual parameters used to generate the data in the simulation studies in Chapters 4 and 5 are described in more detail. The procedures and R-functions mirrored the repeated single-unit trials introduced in the respective chapters closely. However, some small modifications were necessary for practicability reasons. We first describe the data generation for designs with treatment assignment at the unit-level and the parameters used in the simulation study in Chapter 4. In the second part of the chapter, we introduce the data generation procedure for designs with treatment assignment at the cluster-level and the parameters used in the simulation study in Chapter 5.

C.1 Treatment Assignment at the Unit-Level

In this section, the function for data generation for non-randomized designs with treatment assignment at the unit-level is described. Data generation closely followed the repeated single-unit trial introduced generally in Chapter 2 and specifically for the designs considered in the simulation study in Chapter 4 with small modifications for practicability. The complete R-code used to generate the data in the simulation study is given in Listing C.1 on page 259, the parameters names are given in the text in `typewriter`-style if necessary and are labeled according to their names in the R-function.

Data generation started with the generation of clusters and cluster sizes. In order to obtain `c` clusters with a given average cluster size but varying individual sizes, `c` random numbers were generated from a uniform distribution with lower bound `clusmin` and upper bound `clusmax`. These numbers were rounded to the nearest integer and used as cluster sizes $n_{C=c}$. The generation of the cluster sizes was restricted to ensure that the total sample size N was equal to the product of the average cluster size $\overline{n}_C$ and the

number of clusters in the respective condition of the simulation design.

Next, the values of the between-component Z_b were generated by drawing `c` normally distributed random numbers. The expected value μ_{Z_b} (`z.mean`) and the variance σ^2_Z (`z.var.betw`) could be independently specified. After the generation of the between-component, the values of Z were obtained by sampling `n` times from a normal distribution with an expected value of 0 and variance equal to the within-cluster variance of $\sigma^2_{Z_w}$ (`z.var.within`) and adding the so-generated values of the within-component Z_w to Z_b. Hence, the intraclass correlation of the unit-covariate Z was determined by the two parameters σ^2_Z and $\sigma^2_{Z_w}$ in the following way [see also Equation (2.9)]:

$$ICC(Z) = \frac{\sigma^2_{Z_b}}{\sigma^2_{Z_b} + \sigma^2_{Z_w}}. \tag{C.1}$$

Next, the true-outcome variable τ_0 in the control condition and the true-effect variable δ_{10} were generated by specifying separate linear regressions of the true-outcomes τ_0 in the control group and the individual effects δ_{10} on the between-component Z_b, the within-component Z_w and their product:

$$\tau_0 = \gamma_{00} + \gamma_{01} \cdot Z_b + \gamma_{04} \cdot Z_w + \gamma_{05} \cdot Z_b \cdot Z_w + r_{0;C} + \upsilon_{0;U}, \tag{C.2}$$

$$\delta_{10} = \gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w + r_{10;C} + \upsilon_{10;U}. \tag{C.3}$$

The regression weights γ_{00}, γ_{01}, γ_{04}, γ_{05}, γ_{10}, γ_{11}, γ_{14} and γ_{15} could be independently specified. The residuals $r_{0;C}$ and $r_{10;C}$ represented the residual effects of the cluster variable C on the true-outcome variable in the control group and the true-effect variable respectively. They were obtained by sampling `c` times from two independent normal distributions with mean zero and variances $\sigma^2_{r_{0;C}}$ (`tau0clus.var`) and $\sigma^2_{r_{10;C}}$ (`delta10clus.var`) that could be independently specified. The generated values were added to the true-outcomes and true-effects within the corresponding cluster. The residuals $\upsilon_{0;U}$ and $\upsilon_{10;U}$ represented the residual influence of the unit-variable on the true-outcome variable in the control group and the true-effect variable, respectively. They were obtained by sampling `n` times from two independent normal distributions with mean zero and variances $\sigma^2_{\upsilon_{0;U}}$ (`tau0.var`) and $\sigma^2_{\upsilon_{10;U}}$ (`delta10.var`) that could be independently specified.

```
sim.mrt.data = function(n, y.var, z.mean, z.var.betw, z.var.within,
                        tau0.var, delta10.var,
                        tau0clus.var, delta10clus.var,
                        gamma00, gamma01, gamma04, gamma05,
                        gamma10, gamma11, gamma14, gamma15,
                        clus, clus.min, clus.max,
                        g0, g1, g2,
                        x.thresholds=c(0.0), x.slope=1)
    {
       d = F
       while(d == F){
       clus.sizes =
         round(runif(clus-1, min = clus.min - 0.5, max = clus.max + 0.5))
       d = (!(sum(clus.sizes) > (n-clus.min))) &
           (!(sum(clus.sizes) < (n-clus.max)))
    }
    clus.sizes = c(clus.sizes, (n-sum(clus.sizes)))
    cluster.ID = rep(1:clus, times = clus.sizes)
    z.cluster = rnorm(clus, mean = z.mean, sd = sqrt(z.var.betw))
    z.betw = rep(z.cluster, times = clus.sizes)
    z.with = rnorm(n, mean = 0, sd = sqrt(z.var.within))
    z = z.betw + z.with
    res.tau0.cluster = rnorm(clus,mean=0, sd=sqrt(tau0clus.var))
    res.delta10.cluster = rnorm(clus, mean=0, sd=sqrt(delta10clus.var))
    res.tau0.betw = rep(res.tau0.cluster, times = clus.sizes)
    res.delta10.betw = rep(res.delta10.cluster, times = clus.sizes)
    tau0 = gamma00 + gamma01 * z.betw + gamma04 * z.with +
            gamma05 * z.with * z.betw +
            res.tau0.betw +
            rnorm(n, mean=0, sd = sqrt(tau0.var))
    delta10 = gamma10 + gamma11 * z.betw + gamma14 * z.with +
            gamma15 * z.with * z.betw +
            res.delta10.betw +
            rnorm(n, mean=0, sd = sqrt(delta10.var))
    x.info = item.logit(g0 + g1* z.with + g2* (z.betw-z.mean),
        thr=x.thresholds, slope = x.slope)
    x = as.integer(x.info$x)
    y = tau0 + delta10*x + rnorm(n, mean=0, sd=sqrt(y.var))
    cluster.ID = cluster.ID + 100
    d = as.data.frame(cbind(cluster.ID, y, x, z))
    return(d)
}
```

Listing C.1: R-function to generate data according to the repeated single-unit trial for designs with treatment assignment at the unit-level

Table C.1: Varied parameters in the simulation study of the generalized ANCOVA with treatment assignment at the unit-level

Parameter	Subcondition	Values
`clus`		20; 50; 200
`n`		50; 100; 250 · `clus`
`z.var.betw`		$\frac{1}{19}; \frac{1}{9}; \frac{1}{4}; \frac{3}{7}$
`gamma11`		
	$\sigma^2_{Z_b} = \frac{1}{19}$	0; 4.992; 7.487; 9.803
	$\sigma^2_{Z_b} = \frac{1}{9}$	0; 3.444; 5.166; 6.764
	$\sigma^2_{Z_b} = \frac{1}{4}$	0; 2.309; 3.464; 4.536
	$\sigma^2_{Z_b} = \frac{3}{7}$	0; 1.777; 2.665; 3.490
`gamma10`		- `gamma11`
`g1`		0; 0.275, 0.567, 1.07
`g2`		
	$\sigma^2_{Z_b} = \frac{1}{19}$	0; 1.174; 2.4; 4.65
	$\sigma^2_{Z_b} = \frac{1}{9}$	0; 0.787; 1.715; 3.21
	$\sigma^2_{Z_b} = \frac{1}{4}$	0; 0.53; 1.1; 2.145
	$\sigma^2_{Z_b} = \frac{3}{7}$	0; 0.409; 0.862; 1.649

Hence, the average causal effect ACE_{10} was given by the expected value of the true-effect variable δ_{10} given in Equation (C.3):

$$ACE_{10} = \gamma_{10} + \gamma_{11} \cdot E(Z). \tag{C.4}$$

In the next step, the assignment of each unit u to either the treatment or the control condition was determined. This assignment was based on the values of the within-component Z_w and the between-component Z_b. Treatment assignment probabilities were obtained with a logistic assignment function:

$$P(X{=}1 \mid Z, Z_b) = \frac{exp\,(g_0 + g_1 \cdot Z_w + g_2 \cdot [Z_b - E(Z)])}{1 + exp\,(g_0 + g_1 \cdot Z_w + g_2 \cdot [Z_b - E(Z)])}. \tag{C.5}$$

Parameter g_0 determined the average size of treatment and control group, the parameter g_1 determined the stochastic dependency between the within-component Z_w and parameter g_2 determined the dependency of the between-component Z_b. The values of

Z_b were centered around their expected value to simplify the interpretation of the equation and to make the sizes of treatment and control group independent of the choice of the expected value of the unit-covariate Z. Since Z_w and Z_b are uncorrelated by definition, the corresponding partial regression coefficients of the liner logistic regression in Equation (C.5) are equal to the regression weights of the corresponding simple linear logistic regressions. The dichotomous treatment variable X — a value of zero indicated the control group and a value of one indicated the treatment group — was obtained by sampling `n` times from a uniform random distribution with lower bound 0 and upper bound 1 and comparing the resulting values to the unit-specific assignment probabilities obtained from the assignment function specified in Equation (C.5).

The final step of data generation consisted of generating the values of the outcome variable Y according to the following equation:

$$Y = \tau_0 + \delta_{10} \cdot X + \varepsilon_Y. \tag{C.6}$$

The residual ε_Y was generated by drawing from a normal distribution with an expected value of 0 and variance σ_Y^2 (`y.var`).

C.2 Treatment Assignment at the Cluster-Level

In this section, the function for data generation for non-randomized designs with treatment assignment at the cluster-level is described. Data generation closely followed the repeated single-unit trial introduced generally in Chapter 2 and specifically for the designs considered in the simulation study in Chapter 5 with small modifications for practicability. The complete `R`-code used to generate the data in the simulation is given in Listing C.2 on page 263. Again, if necessary, the parameter names are presented in `typewriter`-style in the text and labeled according to their names in the `R`-function.

Data generation started with the generation of clusters and cluster sizes. In order to obtain `c` clusters with a given average cluster size but varying individual sizes, `c` random numbers were generated from a uniform distribution with lower bound `clusmin` and upper bound `clusmax`. These numbers were rounded to the nearest integer and used as cluster sizes $n_{C=c}$. The generation of the cluster sizes was restricted to ensure that the total sample size N was equal to the product of the average cluster size $\bar{n}_C$ and the number of clusters C in the respective condition of the simulation design.

Table C.2: Constant parameters in the simulation study of the generalized ANCOVA with treatment assignment at the unit-level

Parameter	Value
z.mean	1
z.var.within	1
y.var	2
tau0.var	2.25
delta10.var	1.25
tau0clus.var	0.75
delta10clus.var	0.25
g0	0
gamma00	1
gamma01	1
gamma04	2
gamma05	0
gamma14	−0.5
gamma15	0

Next, the values of the between-component Z_b were generated by drawing c normally distributed random numbers. The expected value μ_{Z_b} (z.mean) and the variance σ^2_Z (z.var.betw) could be independently specified. After the generation of the between-component, the values of Z were obtained by sampling n times from a normal distribution with an expected value of 0 and within-cluster variance $\sigma^2_{Z_w}$ (z.var.within) and adding the so-generated values of the within-component Z_w to Z_b. Hence, the intraclass correlation of the unit-covariate Z was determined by the two parameters σ^2_Z and $\sigma^2_{Z_w}$ in the following way [see also Equation (2.9)]:

$$ICC(Z) = \frac{\sigma^2_{Z_b}}{\sigma^2_{Z_b} + \sigma^2_{Z_w}}. \tag{C.7}$$

```
sim.grt.data <- function(n, y.var, z.mean, z.var.betw, z.var.within,
                         tau0.var, delta10.var,
                         tau0clus.var, delta10clus.var,
                         gamma00, gamma01, gamma04, gamma05,
                         gamma10, gamma11, gamma14, gamma15,
                         g0, g1,
                         clus, clus.min, clus.max,
                         x.thresholds=c(0.0), x.slope=1)
{
    d = F
    while(d == F){
      clus.sizes =
        round(runif(clus-1, min = clus.min - 0.5, max = clus.max + 0.5))
      d = (!(sum(clus.sizes) > (n-clus.min))) &
          (!(sum(clus.sizes) < (n-clus.max)))
    }
    clus.sizes = c(clus.sizes, (n-sum(clus.sizes)))
    cluster.ID = rep(1:clus, times = clus.sizes)
    z.cluster = rnorm(clus, mean = z.mean, sd = sqrt(z.var.betw))
    z.betw = rep(z.cluster, times = clus.sizes)
    z.with = rnorm(n, mean = 0, sd = sqrt(z.var.within))
    z = z.betw + z.with
    res.tau0.cluster = rnorm(clus,mean=0, sd=sqrt(tau0clus.var))
    res.delta10.cluster = rnorm(clus, mean=0, sd=sqrt(delta10clus.var))
    res.tau0.betw = rep(res.tau0.cluster, times = clus.sizes)
    res.delta10.betw = rep(res.delta10.cluster, times = clus.sizes)
    tau0 = gamma00 + gamma01 * z.betw + gamma04 * z.with +
            gamma05 * z.with * z.betw +
            res.tau0.betw +
            rnorm(n, mean=0, sd = sqrt(tau0.var))
    delta10 = gamma10 + gamma11 * z.betw + gamma14 * z.with +
            gamma15 * z.with * z.betw +
            res.delta10.betw +
            rnorm(n, mean=0, sd = sqrt(delta10.var))
    x.info = item.logit(g0 + g1*(z.cluster-z.mean),
         thr=x.thresholds, slope = x.slope)
    x = as.integer(x.info$x[cluster.ID])
    y = tau0 + delta10*x + rnorm(n, mean=0, sd=sqrt(y.var))
    cluster.ID = cluster.ID + 100
    d = as.data.frame(cbind(cluster.ID, y, x, z))
    return(d)
}
```

Listing C.2: R-function to generate data according to the repeated single-unit trial for designs with treatment assignment at the cluster-level

Table C.3: Varied parameters in the simulation study of the generalized ANCOVA with treatment assignment at the cluster-level

Parameter	Subcondition	Values
`clus`		5; 10; 25; 50
`n`		20; 50; 100; 200 · `clus`
`z.var.betw`		$\frac{1}{19}$; $\frac{1}{9}$; $\frac{1}{4}$; $\frac{3}{7}$
`gamma11`		
	$\sigma^2_{Z_b} = \frac{1}{19}$	0; 2.298; 3.435, 4.992; 7.487; 9.803
	$\sigma^2_{Z_b} = \frac{1}{9}$	0; 1.654; 2.370; 3.444; 5.166; 6.764
	$\sigma^2_{Z_b} = \frac{1}{4}$	0; 1.109; 1.589; 2.309; 3.464; 4.536
	$\sigma^2_{Z_b} = \frac{3}{7}$	0; 0.854; 1.223; 1.777; 2.665; 3.490
`gamma10`		- `gamma11`
`g1`		
	$\sigma^2_{Z_b} = \frac{1}{19}$	0; 1.174; 2.4; 4.65
	$\sigma^2_{Z_b} = \frac{1}{9}$	0; 0.787; 1.715; 3.21
	$\sigma^2_{Z_b} = \frac{1}{4}$	0; 0.53; 1.1; 2.145
	$\sigma^2_{Z_b} = \frac{3}{7}$	0; 0.409; 0.862; 1.649

Next, the true-outcome variable τ_0 in the control condition and true-effect variable δ_{10} were generated by specifying separate linear regressions of the true-outcome variable in the control group τ_0 and the true-effect variable δ_{10} on the between-component Z_b, the within-component Z_w and their product:

$$\tau_0 = \gamma_{00} + \gamma_{01} \cdot Z_b + \gamma_{04} \cdot Z_w + \gamma_{05} \cdot Z_b \cdot Z_w + r_{0;C} + \upsilon_{0;U}, \tag{C.8}$$

$$\delta_{10} = \gamma_{10} + \gamma_{11} \cdot Z_b + \gamma_{14} \cdot Z_w + \gamma_{15} \cdot Z_b \cdot Z_w + r_{10;C} + \upsilon_{10;U}. \tag{C.9}$$

Again, the regression weights γ_{00}, γ_{01}, γ_{04}, γ_{05}, γ_{10}, γ_{11}, γ_{14} and γ_{15} could be independently specified. The residuals $r_{0;C}$ and $r_{10;C}$ represented the residual effects of the cluster variable C on the true-outcomes in the control group and the true-effects respectively. They were obtained by sampling `c` times from two independent normal distributions with mean zero and variances $\sigma^2_{r_{0;C}}$ (`tau0clus.var`) and $\sigma^2_{r_{10;C}}$ (`delta10clus.var`) that could be independently specified. The generated values were added to the true-outcomes and true-effects within the corresponding cluster. The residuals $\upsilon_{0;U}$ and $\upsilon_{10;U}$ represented the residual influence of the unit-variable on the true-outcome vari-

Table C.4: Constant parameters in the simulation study of the generalized ANCOVA with treatment assignment at the cluster-level

Parameter	Value
z.mean	1
z.var.within	1
y.var	2
tau0.var	2.25
delta10.var	1.25
tau0clus.var	0.75
delta10clus.var	0.25
g0	0
gamma00	1
gamma01	1
gamma04	2
gamma05	0
gamma14	−0.5
gamma15	0

able in the control group and the true-effect variable, respectively. They were obtained by sampling n times from two independent normal distributions with mean zero and variances $\sigma^2_{\upsilon_{0;U}}$ (tau0.var) and $\sigma^2_{\upsilon_{10;U}}$ (delta10.var) that could be independently specified.

The average causal effect ACE_{10} was thus given by the expected value of the true-effect variable δ_{10} as given in Equation (C.9):

$$ACE_{10} = \gamma_{10} + \gamma_{11} \cdot E(Z). \tag{C.10}$$

In the next step, the assignment of each cluster c to either the treatment or the control condition was determined. The assignment probabilities depended on the values of the between-component Z_b. Treatment assignment probabilities were obtained from a logistic assignment function:

$$P(X{=}1 \mid V, Z_b) = \frac{exp\,(g_0 + g_1 \cdot [Z_b - E(Z)])}{1 + exp\,(g_0 + g_1 \cdot [Z_b - E(Z)])} \tag{C.11}$$

The parameter g_0 determines the average number of clusters assigned to the treatment

condition; the parameter g_1 determines the stochastic dependence between X and and the between-component Z_b (g_1). The values of Z_b were centered around their expected value to simplify the interpretation of the equation and to make the sizes of treatment and control group independent of the choice of the expected value of the unit-covariate Z. The dichotomous treatment variable X — a value of zero indicated the control group and a value of one indicated the treatment group — was obtained by sampling `c` times from a uniform random distribution with lower bound 0 and upper bound 1 and comparing the resulting values with the unit-specific assignment probabilities obtained from the assignment function specified in Equation (C.11).

In the final step of the data generation process, the outcome variable Y was generated according to the following equation:

$$Y = \tau_0 + \delta_{10} \cdot X + \varepsilon_Y. \tag{C.12}$$

The residual ε_Y was generated from a normal distribution with expected value of 0 and variance σ_Y^2 (`y.var`).

D Dependent Variables in the Simulations

Convergence. The convergence rate was computed as the number of converged solutions relative to the number of replications per cell of the simulation design. Convergence rates were calculated separately for each method.

Bias of *ACE*-Estimator. Since the true average causal effect was zero in all conditions of the simulations, the mean bias of the estimator of the average causal effect was simply defined as the mean of the parameter estimate over the replications in a simulation condition:

$$MB(\widehat{ACE}_{10}) = \overline{\widehat{ACE}_{10}}. \tag{D.1}$$

In accordance with the recommendations by Boomsma and Hoogland (2001), an absolute bias of $|MB| \leq 0.025$ was deemed acceptable.

Efficiency of *ACE*-Estimator. The mean squared error (*MSE*) was used to compare the relative efficiency of the estimation method. Since the *ACE* was equal to zero in all experimental conditions, the *MSE* of the *ACE*-estimator in a condition of the simulation design is given by the following equation:

$$MSE(\widehat{ACE}_{10}) = \frac{1}{R}\sum_{i=0}^{R}(\widehat{ACE}_{10i})^2, \tag{D.2}$$

where R is the number of replications per cell.

Relative Bias of Standard Error. The bias of the standard errors were evaluated using the mean relative bias (*MRB*, Boomsma & Hoogland, 2001). The *MRB* of the

standard error of the average causal effect estimator was defined as follows:

$$MRB\left[\widehat{SE}(\widehat{ACE}_{10})\right] = \frac{\overline{\widehat{SE}}(\widehat{ACE}_{10}) - SD(\widehat{ACE}_{10})}{SD(\widehat{ACE}_{10})} \tag{D.3}$$

where $\overline{\widehat{SE}}(\widehat{ACE}_{10})$ is the mean of the estimated standard errors and $SD(\widehat{ACE}_{10})$ is the observed standard deviation of the parameter estimates. Both values were calculated over $N_R = 1000$ replications per cell in the present simulation. The relative bias measure can be interpreted as the percentage of bias in a parameter estimator relative to the true value of the parameter. It is independent of the true parameter value and can thus be used to compare the biasedness of standard error estimators whose true values substantially decrease with larger sample sizes.

Type-1-Error Rate. The adherence to the nominal α-level, i.e., the actual type-1-error rate, when testing the H_0: $ACE_{10} = 0$ was also analyzed. The *rejection frequency* (RF) over $N_R = 1000$ replications, based upon the number of significant results for two-sided-tests of the H_0: $ACE_{10} = 0$ using the respective reference distributions for the test statistics was used to determine the control the nominal significance-level for every method in each cell of the experimental design. The RFs were calculated for three commonly used significance levels $\alpha_1 = 0.1$, $\alpha_2 = 0.05$ and $\alpha_3 = 0.01$.

Given N_R and α, the rejection frequency follows a binomial distribution, i.e., $RF \sim Bin(NR, \alpha)$. Hence, confidence intervals can be calculated, that contain the empirical value of the RF with a certain probability when the nominal α-level is controlled by the test. These intervals are used as criteria for acceptable behavior of the z-score (Boomsma & Hoogland, 2001). The 95%-prediction intervals for experimental cells were constructed for every significance level and resulted in the following intervals: $CI(\alpha{=}0.1) = [82, 119]$, $CI(\alpha{=}0.05) = [37, 64]$ and $CI(\alpha{=}0.01) = [4, 17]$, i.e., when testing at the respective significance levels, only 5% of the observed RF should fall outside the prediction intervals, if the α-level is properly controlled by the test statistic.

E Contents of the Accompanying CD

The accompanying CD contains the raw results of the simulation studies presented in Chapters 4 and 5 as text files that can be imported into any statistical package. Additionally, graphics for all statistical methods and dependent measures are given that are similar to the graphics presented in the aforementioned chapters.

The file system of the CD is structured as follows, file names are self-explanatory:

- `Unit`: Contains the results of the simulation for treatment assignment at the unit-level.
  - `data`: Contains separate raw data files for each dependent measure.
  - `pics`: Contains the graphic files for ...
    - `convergence`: ... average convergence rates,
    - `bias_ace`: ... mean bias of the *ACE*-estimators,
    - `bias_se`: ... mean relative bias of the standard error estimators,
    - `alpha`: ... type-1-error rates.
- `Cluster`: Contains the results of the simulation for treatment assignment at the cluster-level.
  - `data`: Contains separate raw data files for each dependent measure.
  - `pics`: Contains the graphic files for ...
    - `convergence`: ... average convergence rates,
    - `bias_ace`: ... mean bias of the *ACE*-estimators,
    - `bias_se`: ... mean relative bias of the standard error estimators,
    - `alpha`: ... type-1-error rates.

References

Aiken, L. S., & West, S. G. (1996). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Algina, J. (1999). A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *34*, 493-504.

Alker, H. R. j. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.), *Quantitative ecological analysis in the social sciences* (pp. 69–86). Cambridge, MA: M.I.T. Press.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Asparouhov, T., & Muthén, B. O. (2004). *Full-information-maximum-likelihood estimation of general two-level latent variable models.* Unpublished manuscript.

Asparouhov, T., & Muthén, B. O. (2006). *Constructing covariates in multilevel regression* (Tech. Rep.). Los Angeles, CA: Muthén and Muthén: Mplus Web Notes: No. 11. Retrieved January 9, 2007 from www.statmodel.com.

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or Type I Errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, *73*, 924-935.

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes [Computer software and manual]. Retrieved November 27, 2008 from http://lme4.r-forge.r-project.org/.

Bauer, D. J., & Cai, L. (2008). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Eduational and Behavioral Statistics*.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163.

Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based in-

terventions when control participants are ungrouped. *Multivariate Behavioral Research*, *43*, 210-236.

Bauer, H. (1981). *Probability theory and elements of measure theory* (2nd ed.). London: Acadamic Press.

Bingenheimer, J. B., & Raudenbush, S. W. (2004). Statistical and substantive inferences in public health: Issues in the application of multilevel models. *Annual Review of Public Health*, *25*, 53-77.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bloom, H. S., Bos, J. M., & Lee, S. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*, 445–469.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30-59.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In *Structural Equation Modeling: Present and Future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago: Scientific Software International.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton-Mifflin.

Chen, X. (2006). The adjustment of random baseline measurements in treatment effect estimation. *Journal of Statistical Planning and Inference*, *136*, 4161-4175.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, designs and analysis*. Stanford, CA: Stanford Evaluation Consortium.

Croon, M. A., & van Veldhoven, M. J. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45-57.

Curtin, T. R., Ingels, S. J., Wu, S., Heuer, R., & Owings, J. (2002). *National Education Longitudinal Study of 1988: Base-year to fourth follow-up data file user's manual*. Washington, D. C.: National Center for Education Statistics.

Diez Roux, A. V. (2004). Estimating neighborhood health effects: The challenges of causal inference in a complex world. *Social Science & Medicine*, *58*, 1953-1960.

Donner, A., & Klar, N. (2000). *Cluster randomization trials in health research*. London: Arnold.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, *20*, 115-147.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Thousand Oaks, CA: Sage Publications.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Feng, Z., Diehr, P., Peterson, A., & McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health*, *22*, 167-187.

Fiege, C. (2007). *Faire Vergleiche in Schulleistungsuntersuchungen und ihre kausaltheoretische Grundlage [Fair comparisons in school effectiveness studies and their causal interpretation]*. Unpublished diploma thesis, Friedrich-Schiller-Universität Jena, Germany.

Fisicaro, S. A., & Tisak, J. (1994). A theoretical note on the stochastics of moderated multiple regression. *Educational and Psychological Measurement*, *54*, 32–41.

Flory, F. (2008). *Average treatment effects in structural equation models with interactions between treatment and manifest or latent covariates*. Unpublished doctoral dissertation, Friedrich-Schiller-Universität Jena, Germany.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Frangakis, C. E., Rubin, D. B., & Zhou, X. (2002). Clustered encouragement designs with individual noncompliance: Bayesian inference with randomization, and application to advance directive forms. *Biostatistics*, *3*, 147–164.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516-524.

Gelman, A., & Hill, J. L. (2007). *Data analysis using regression and multilevel / hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Pardoe, I. (2007). Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, *37*, 23–51.

Gitelman, A. I. (2005). Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, *30*, 397–412.

Goedman, R., Grothendieck, G., Højsgaard, S., & Pinkus, A. (2007). *Ryacas: R interface to the yacas computer algebra system.* [Computer software and manual]. Retrieved August 8, 2008 from http://ryacas.googlecode.com.

Goldstein, H. (1999). *Multilevel statistical models*. London: Edvard Arnold.

Goldstein, H., Browne, W., & Rasbash, I. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*, 223–231.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine*, *11*, 1209-1223.

Grilli, L., & Rampichini, C. (2008). *Measurement error in multilevel models with sample cluster means.* Unpublished manuscript. Unpublished manuscript.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. Data from Health Survey for England 1994. *American Journal of Epidemiology*, *149*, 876-883.

Halloran, M. E., & Struchiner, C. J. (1995). Causal inference in infectious diseases. *Epidemiology*, *6*, 142-151.

Hayes, R., & Bennet, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, *28*, 319–326.

Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151-179.

Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate answers from randomized experiments. *Psychological Methods*, *1*, 154-169.

Helgeson, V. S., Cohen, S., Schulz, R., & Yasko, J. (1999). Education and peer dis-

cussion group interventions and adjustment to breast cancer. *Archives of General Psychiatry*, *56*, 340-347.

Hinkelmann, K., & Kempthorne, O. (2005). *Design and analysis of experiments*. Hoboken, NJ: Wiley.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, *3–4*, 259–278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945-960.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*, 901-910.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*, 333–362.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Huber, P. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (p. 221-233). Berkeley, CA: University of California Press.

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncomplicance. *Psychological Methods*, *13*, 1-18.

Jöreskog, K. G., & Sörbom, D. (1996 - 2001). *LISREL 8: User's reference guide* (2nd ed.). Lincolnwood, IL: Scientific Software International.

Kang, J. D. Y., & Schafer, J. L. (2006). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean for incomplete data. *Statistical Science*, *22*, 523–539.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology*, *83*, 126–137.

Kim, J.-S., & Frees, E. W. (2006). Omitted variables in multilevel models. *Psychome-*

trika, *71*, 659-690.

Kim, J.-S., & Frees, E. W. (2007). Multilevel modeling with correlated effects. *Psychometrika*, *72*, 505–533.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kolmogorov, A. N. (1933/1956). *Foundations of the theory of probability*. New York: Chelsea.

Konstantopoulos, S. (2008). Computing power of tests of the variance of treatment effects in designs with two levels of nesting. *Multivariate Behavioral Research*, *43*, 327-352.

Korendijk, E. J., Maas, C. J., Moerbeek, M., & Van der Heijden, P. G. (2008). The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. *Methodology*, *4*, 67-72.

Kozlowski, S. W. J., & Klein, K. J. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco: Jossey-Bass.

Kreft, I. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*, 1-21.

Kröhne, U. (2007). *SimRobot [Computer software].* Department of Psychology, Friedrich-Schiller-Universität Jena, Germany.

Kröhne, U. (2009). *Estimation of average causal effects in quasi-experimental designs: Nonlinear constraints in structural equation models*. Unpublished doctoral dissertation, Friedrich-Schiller-Universität Jena, Germany.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 248-277.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical model* (5 ed.). Bosten: McGraw-Hill.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Lechner, M. (2001). *A note on the common support problem in applied evaluation studies* (Tech. Rep.). University of St. Gallen Economics, Disc. Paper 2001-01.

Loève, M. (1977). *Probability theory I* (4th ed.). New York: Springer.

Loève, M. (1978). *Probability theory II* (4th ed.). New York: Springer.

Longford, N. T. (1995). Hierarchical models and social sciences. *Journal of Educa-*

tional and Behavioral Statistics, *20*, 205-209.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum.

Manski, C. (1995). *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press.

Marsh, H. W., Hau, K., & Craven, R. G. (2004). The big-fish-little-pond effect stands up to scrutiny. *American Psychologist*, *59*, 268–271.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101.

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. *Journal of Educational and Behavioral Statistics*, *33*, 41-61.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25*, 271-284.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental designs for multilevel model with covariates. *Communications in Statistics - Theory and Methods*, *30*, 2683–2697.

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*, 341-350.

Moerbeek, M., van Breukelen, G. J. P., Berger, M. P. F., & Ausems, M. (2003). Optimal sample sizes in experimental designs with individuals nested within clusters. *Understanding Statistics*, *2*, 151-175.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New York:

Wiley.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference. Methods and principles for social research*. New York: Cambridge University Press.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Murray, D. M. (2001). Statistical models appropriate for designs often used in group-randomized trials. *Statistics in Medicine*, *20*, 1373-1385.

Murray, D. M., Van Horn, M. L., Hawkins, J. D., & Arthur, M. W. (2006). Analysis strategies for a community trial to reduce adolescent ATOD use: A comparison of random coefficient and ANOVA/ANCOVA models. *Contemporary Clinical Trials*, *27*, 188-206.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, *94*, 423-432.

Muthén, B. O. (1998-2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117.

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–166). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B. O., Khoo, S.-T., & Gustafsson, J.-E. (1997). *Multilevel latent variable modeling in multiple populations* (Tech. Rep.). Graduate School of Education & Information Studies, University of California, Los Angeles.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5 ed.). Los Angeles, CA: Muthén & Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.

Nagengast, B. (2006). *Standard errors of ACE estimates: Comparing adjusted group means against the adjusted grand mean. A simulation study*. Unpublished diploma thesis, Friedrich-Schiller-Universität Jena, Germany.

National Center for Education Statistics. (2001). *User's manual for the ECLS-K base year public-use data files and electronic codebook*. Washington, D.C.: National

Center for Education Statistics.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*, 465–480.

Oakes, J. M. (2004). The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology. *Social Science & Medicine*, *58*, 1929–1952.

Partchev, I. (2007). *lace: Estimates and tests for average causal effects via LISREL and Mplus.* [Computer software and manual]. Retrieved March 1, 2008, from www.statlite.com. Retrieved march 1, 2008, from www.statlite.com.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., & the R Core team. (2008). *nlme: Linear and nonlinear mixed effects models.* [Computer software and manual] R package version 3.1-90.

Pituch, K. A. (2001). Using multilevel modeling in large-scale planned variation educational experiments: Improving understanding of intervention effects. *Journal of Experimental Education*, *69*, 347–373.

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, *41*, 367-400.

Pituch, K. A., Whittaker, T. A., & Stapleton, L. M. (2005). A comparison of methods to test for mediation in multisite experiments. *Multivariate Behavioral Research*, *40*, 1–23.

Plewis, I., & Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Educational Research and Evaluation*, *4*, 13–26.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria.

Rao, C. R. (1973). *Linear statistical inference and its applications*. London: Wiley.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2005). *A user's guide to MLwiN version 2.0.* [Computer software and manual]. Retrieved November 26, 2008,

from http://www.mlwin.com/download/userman_2005.pdf.

Raudenbush, S. W. (1995). Reexamining, reaffirming, and improving application of hierarchical models. *Journal of Educational and Behavioral Statistics*, *20*, 210-220.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173-185.

Raudenbush, S. W. (2003). *Designing field trials of educational innovations* (Tech. Rep.). Ann Arbor, MI: University of Michigan, Department of Sociology.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., Hong, G., & Rowan, B. (2006). Studying the causal effects of instruction with application to primary-school mathematics. In J. Ross, G. Bohrnstedt, & F. Hemphill (Eds.), *Instructional and performance of high poverty schooling*. Washington, D.C.: National Council for Educational Statistics.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199-213.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29.

Raudenbush, S. W., & Sampson, R. (1999). Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods & Research*, *28*, 123-153.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, *11*, 621-637.

Rechner, A. C., & Schaalje, G. B. (2007). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley.

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*, 152-162.

Robins, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, *40*, 139–161.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Statistical Science*, 6–10.

Robins, J. M., Hérnan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550–560.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for the NELS:1988 base-year through second follow-up*. Washington, D.C.: National Center for Education Statistics.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*, 286–304.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, *102*, 191-200.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33–38.

Rubin, D. B. (1973). The use of matching and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 185–203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26.

Rubin, D. B. (1978). Bayesian-inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1986). Statistics and causal inference - which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Rubin, D. B. (1990). Formal modes of statistical-inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279–292.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *3–4*, 169–188.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *103*, 1350-1353.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*, 103–116.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, *69*, 682–689.

Schafer, J. L., & Kang, J. D. Y. (2008). Average causal effects from observational studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 259–280). Thousand Oaks, CA: Sage.

Shadish, W. R. (2000). The empirical program of quasi-experimentation. In L. Bickman (Ed.), *Research design: Donald Campbell's legacy* (pp. 13–35). Thousand Oaks, CA: Sage.

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods*, *7*, 3-18.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random to nonrandom assignment. *Journal of the American Statistical Association*, *103*,

1334-1343.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., & Heinsman, D. T. (1997). Experiments versus quasi-experiments: Do they yield the same answer? *NIDA Research Monograph*, *170*, 147–164.

Shadish, W. R., & Luellen, J. K. (2005). Quasi-experimental designs. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1641–1644). Chichester: Wiley.

Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. Bootzin & P. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.

Shieh, G. (2006). Exact interval estimation, power calculation, and sample size determination in normal correlational analysis. *Psychometrika*, *71*, 529–540.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*, 137–167.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, *101*, 1398-1407.

Sobel, M. E. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, *33*, 230–251.

Spybrook, J., Raudenbush, S. W., Liu, X., Congdon, R., & Martínez, A. (2008). *Optimal design for longitudinal and multilevel research: Documentation for the "optimal design" software* (Tech. Rep.). Retrieved November 26, 2008, from http://sitemaker.umich.edu/group-based/optimal_design_software.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger

(Eds.), *What if there were no significance tests?* (p. 221-257). Mahwah, NJ: Erlbaum.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle [Theory of causal regression models]*. Stuttgart: Fischer.

Steyer, R. (2002). *Wahrscheinlichkeit und Regression [Probability and regression]*. Berlin: Springer.

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online*, *5*, 55–87.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C., & Buhl, T. (2000). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online*, *5*, 39–71.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*, *7*, 41–68.

Steyer, R., & Partchev, I. (2008). *EffectLite for Mplus: A program for the uni- and multivariate analysis of unconditional, conditional and average mean differences between groups [Computer software and manual].* Retrieved May 5, 2008, from www.statlite.com.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (2009). *Probability and causality*. Heidelberg: Springer. (Version: June 26, 2009)

Subramanian, S. V. (2004). The relevance of multilevel statistical methods for identifying causal neighborhood effects. *Social Science & Medicine*, *58*, 1961-1967.

Turner, R., Thompson, S., & Spiegelhalter, D. (2005). Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials*, *2*, 108–118.

Turpin, R. S., & Sinacore, J. M. (Eds.). (1991). *Multisite evaluations*. San Francisco: Jossey-Bass.

Ukoumunne, C. C., Gulliford, M. C., Chinn, S., Sterne, J. A. C., & Burney, P. G. J. (1999). Methods for evaluating area-wide and organisation based interventions in health and health care: A systematic review. *Health Technology Assessment*, *3*, 1-99.

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*,

368-392.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, *27*, 1934–1943.

Varnell, S. P., Murray, D. M., Janega, J. B., & Blitstein, J. S. (2004). Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health*, *94*, 393-399.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4 ed.). New York: Springer.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Walsh, J. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics*, *18*, 88–96.

Wegscheider, K. (2004). Methodische Anforderungen an Einrichtungsvergleiche (Profiling) im Gesundheitswesen [Methodical requirements for institutional comparisons in public health service]. *Zeitschrift für ärztliche Fortbildung und Qualität im Gesundheitswesen*, *98*, 647–654.

Werner, J. (1997). *Lineare Statistik: Das Allgemeine Lineare Modell [Linear statistics: The general linear model]*. Weinheim: Beltz.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continous variable interactions. *Journal of Personality*, *64*, 1–48.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1-25.

Ehrenwörtliche Erklärung

Die Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität in der geltenden Fassung ist mir bekannt.

Ich habe diese Dissertation selbst angefertigt und dabei insbesondere die Hilfe eines Promotionsberaters nicht in Anspruch genommen. Alle von mir benutzten Quellen und Hilfsmittel habe ich kenntlich gemacht und an den entsprechenden Stellen angegeben.

Ulf Kröhne hat mich bei der technischen Durchführung und Programmierung der Simulationsstudien mit `SimRobot` unentgeltlich unterstützt. Christiane Fiege und Norman Rose haben unentgeltlich Vorabversionen des Manuskripts gelesen und mich auf Fehler und Inkonsistenzen aufmerksam gemacht. Remo Kamm hat mich im Rahmen seiner Tätigkeit als studentische Hilfskraft am Lehrstuhl für Methodenlehre und Evaluationsforschung bei der Zusammenstellung des Literaturverzeichnis entgeltlich unterstützt. Darüber hinaus habe ich keine weitere entgeltliche oder unentgeltliche Unterstützung bei der Auswahl und Auswertung des Materials und bei der Herstellung des Manuskripts erhalten.

Darüber hinaus haben Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe diese Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

Ich habe weder die gleiche noch eine in wesentlichen Teilen ähnliche noch eine andere Arbeit bei einer anderen Hochschule oder Fakultät als Dissertation eingereicht.

Ich versichere, dass die oben gemachten Angaben nach meinem besten Wissen der Wahrheit entsprechen und ich nichts verschwiegen habe.

Jena, den _____ _____

Benjamin Nagengast

Lebenslauf

Benjamin Nagengast

Geboren am 14.05.1979 in Mainz

Familienstand: ledig

1985 - 1989	Phillipp-Keim-Schule, Hofheim-Diedenbergen (Grundschule)
1989 - 1991	Heiligenstockschule, Hofheim-Marxheim (Förderstufe)
1991 - 1995	Gesamtschule-Am-Rosenberg, Hofheim-Marxheim
1995 - 1998	Main-Taunus-Schule, Hofheim
15.06.1998	Abitur
1998 - 1999	Zivildienst
1999 - 2002	Psychologiestudium, Ruprecht-Karls-Universität Heidelberg
14.12.2001	Vordiplom in Psychologie, Ruprecht-Karls-Universität Heidelberg
2002 - 2003	Auslandsstudium, The Ohio State University, Columbus, Ohio, USA
2003 - 2006	Psychologiestudium, Friedrich-Schiller-Universität Jena
06.04.2006	Diplom in Psychologie, Friedrich-Schiller-Universität Jena
2006 - 2009	Wissenschaftlicher Mitarbeiter und Promotionsstudent, Institut für Psychologie, Lehrstuhl für Methodenlehre und Evaluationsforschung, Friedrich-Schiller-Universität Jena

Jena, 26.02.2009